Universidade de Lisboa
**Faculdade de Ciências**

Departamento de Estatística e Investigação Operacional


# SEGURO DE SAÚDE: CUSTOS DE AMBULATÓRIO - MODELIZAÇÃO LINEAR GENERALIZADA

Maria do Carmo de Ornelas R. Marques Bandeira

DISSERTAÇÃO

**Mestrado em Estatística**

2013

Universidade de Lisboa
**Faculdade de Ciências**

Departamento de Estatística e Investigação Operacional


# SEGURO DE SAÚDE: CUSTOS DE AMBULATÓRIO - MODELIZAÇÃO LINEAR GENERALIZADA

Maria do Carmo de Ornelas R. Marques Bandeira

DISSERTAÇÃO
**Mestrado em Estatística**

Orientadora:

Professora Doutora Maria Isabel Calisto Frade Barão

2013

Seguro De Saúde: Custos De Ambulatório - Modelização Linear Generalizada

# Agradecimentos

A realização desta tese, que foi para mim um grande desafio, só foi possível graças ao apoio e colaboração da equipa de gestão da Multicare, professores, colegas familiares e amigos. Embora seja impossível agradecer da forma devida a todas as pessoas que o mereciam, não posso deixar de expressar os meus sinceros agradecimentos, em particular:

À Professora Isabel Barão, minha orientadora, pelas sugestões e correcções feitas durante a orientação e em especial pela disponibilidade, encorajamento e apoio nos momentos de maior pressão.

Ao Professor João Gomes e À Prof.ª Teresa Alpuim pela sua ajuda, atenção e disponibilidade manifestada.

Ao Conselho de Administração da Multicare, em particular ao Sr. Dr. Nunes Coelho pelo seu excepcional apoio e à Sr.ª Dr.ª Maria João Sales Luís, pelo entusiasmo em todos sucessos que fui obtendo no decorrer deste mestrado.

A todo o grupo do Gabinete de Actuariado e Controlo pelo carinho e boa disposição com que sempre se apresentam para responder aos desafios que nos vão sendo colocados e, em particular, ao Pedro Marcelino pela imprescindível colaboração, à Marli Amorim e à Joana Fernandes pela dedicação e pela disponibilidade para os debates estatísticos que fomos fazendo ao longo deste percurso e à Marta Sardinha pela sua amizade e colaboração.

À minha família, e em especial aos meus pais, Josefina e Carlos, pelo amor incondicional, compreensão e ternura que sempre me dedicaram, à minha tia Mimi, que sempre nos vai tentando empurar para a vida académica, a que ela tanto se dedicou, aos oito magníficos: meus filhos André, Maria e Tomás e meus sobrinhos Gonçalo, Filipa, Manuel Maria, Francisco e Marta e aos meus irmãos, Rita e João, e cunhados, Gi e Manel, que tanta animação e força trazem à minha vida.

A todos os meus amigos, por perdoarem as minhas inúmeras ausências e por sempre me apoiarem. Em particular, às minhas amigas, Joana Carrilho, Maria João Esteves e Cristina Cardoso.

A todos, muito obrigado.

Seguro De Saúde: Custos De Ambulatório - Modelização Linear Generalizada


## Resumo

A actual forma de cálculo de preço dos seguros de saúde em Portugal distingue, para além do nível de coberturas que integram cada apólice, os riscos consoante se se trata de seguros individuais ou seguros de grupo e consoante as idades das pessoas seguras. Tem-se assistido, nos últimos anos, a uma grande pressão sobre os prémios praticados, o que tem exigido das Seguradoras um conhecimento cada vez maior do risco que contratam. Assim é emergente a necessidade de conhecer cada vez melhor o risco *a priori*.

Este estudo incidiu sobre a cobertura de Ambulatório. Para isso, dado que as nossas variáveis eram muitas e o detalhe muito grande, recorreu-se à análise de clusters para a redução do número de variáveis a incluír no modelo.

A metodologia aqui utilizada foram os Modelos de Regressão, em que se experimentou em simultâneo os GLM's: Gamma, Logit, Probit e Normal com variável endógena transformada.

Com o objectivo de conhecer o custo do risco achou-se apropriado combinar os modelos Logit/Probit para a ocorrência de sinistros (utilização do seguro) e os restantes (Gamma e Gaussiano com variável endógena transformada) para a severidade.

Chegou-se a um conjunto de variáveis com melhor poder explicativo, o que alargou o leque de variáveis habitualmente utilizadas na tarifação desta cobertura, apesar de a variabilidade explicada ter sido apenas de 18%



**Palavras-chave:** GLM, Logit, Gamma, *Clusters*, seguro de saúde, tarifação



## Abstract

Nowadays,the health insurance pricing in Portugal distinguishes, beyond the level of coverage within each policy, the risks depending on whether it is individual insurance or group insurance and depending on the ages of the insured persons. There has been, in recent years, a lot of pressure on premiums charged, which has required a better knowledge of the Insurers on the risk they sell. Therefore, there's an emerging need to know, even better, the *a priori* risk.

This study approached the outpatients coverage. For this, since our variables were many and thoroughly detail, we used the cluster analysis to reduce the number of variables to be included in the model.

The methodology used was the Regression Models, in which we tried simultaneously the GLM's :Gamma, Logit, Probit and Normal with transformed endogenous variable.

In order to meet the cost of risk it was found to be appropriate to combine the Logit / Probit models for the occurrence of events (insurance usage), and the remaining models (Gamma and Gaussian transformed with endogenous variable) for severity.

A set of significant explanatory variables was reached, which extended the range of variables usually used in the pricing of this cover, though it explained only 18% of the variability.



**Keywords:** GLM, Logit, Gamma, Clusters, Health insurance, pricing

# Índice

# Índice de Figuras

# Índice de Tabelas



# Lista de Abreviaturas e Notações

| | |
|---|---|
| $N(\mu,\sigma^2)$ | Distribuição Normal de valor médio $\mu$ e variância $\sigma^2$ |
| $N_k(U,\Sigma)$ | Distribuição Normal Multivariada com parâmetros dados pelo vector U (vector dos valores médios) e $\Sigma$ (matriz de covariâncias), em que k é a ordem da matriz $\Sigma$ |
| $\chi^2_n$ | Distribuição Qui-Quadrado com n graus de liberdade |
| $q_{1-\alpha}$ | Quantil de probabilidade 1 – α da distribuição Normal(0,1) |
| $\chi^{1-\alpha}_n$ | Quantil de probabilidade 1 – α da distribuição Qui-Quadrado com n graus de liberdade |
| $I_n$ | Matriz identidade de ordem n |
| **T.L.C.** | Teorema do Limite Central |
| **i.i.d.** | Independentes e identicamente distribuídas |
| **f.m.p.** | Função massa de probabilidade |
| **f.d.p.** | Função densidade de probabilidade |

# 1. Introdução

## 1.1. Motivação e Objetivos

O desenho técnico de um seguro de saúde constrói-se a partir da definição das coberturas que se pretendem disponibilizar, sendo que, nalguns produtos, essas coberturas poderão ser coberturas opcionais. Assim, sempre que se calcula o prémio de um produto de seguro, deve-se fazê-lo cobertura a cobertura.

O Seguro assenta numa base de mutualidade, isto é, da partilha: todos pagam para suportar as despesas dos que precisam.

Quando olhamos para a atividade seguradora verificamos que existem, no entanto, classes tarifárias; isto é, nem todos os clientes pagam o mesmo para beneficiar de idêntica cobertura de risco. Várias razões existem para essa diferenciação: razões técnicas – o objeto do seguro é o risco e não o facto consumado – e razões comerciais – quem se quer ter como clientes? Uma vez que o nível de exposição ao risco define o preço, deve-se, por um lado, cobrar diferentes prémios sempre que os indivíduos representarem diferentes apetências ao risco, por outro, o mesmo prémio quando se tratar da mesma classe de risco.

Consoante a frequência e a variabilidade com que os fatores de risco se apresentam, estes devem ser utilizados como variáveis definidoras das classes de risco ou de sobreprémios. Este estudo incidirá sobre as que apresentam maior frequência e variabilidade – definidoras das classes de tarifação – deixando de parte as que, em oposição, por serem mais raras serão apenas variáveis justificativas de agravamento do prémio.

A tarifação do Seguro de Saúde suporta-se tradicionalmente na idade da Pessoa Segura e no tipo de contrato de transferência de risco em que se consubstancia:

- ✓ Individual ou coletivo (designado habitualmente por seguro de Grupo);
- ✓ Coberturas garantidas (Internamento, Ambulatório, Estomatologia, Medicamentos, Próteses e Ortóteses ou outras)
- ✓ Limite dos riscos segurados (Capital máximo coberto, percentagens de comparticipação da Seguradora, franquias a cargo da Pessoa Segura,…)

Efetivamente, em termos de características pessoais que possam influenciar a tipologia de risco que cada Individuo ou grupo de indivíduos representa, apenas a idade e o género são analisados, sendo que este último deixou de poder ser utilizado a partir de 21 de Dezembro de 2012[1].

Como atuária e após 24 de anos de atividade seguradora, sendo os últimos 6 dedicados em exclusivo ao ramo Doença e tendo em consideração a importância crescente deste ramo de seguros na vida da sociedade portuguesa, cresce a necessidade imperiosa de identificar quais as variáveis explicativas, essencialmente características pessoais, que nos podem trazer informação acrescentada ao conhecimento do risco *a priori*.

[1] **Orientações sobre a aplicação ao setor dos seguros da Diretiva 2004/113/CE do Conselho, à luz do acórdão do Tribunal de Justiça da União Europeia no Processo C-236/09 (Test-Achats)**

Assim, este projeto tem como objetivo identificar as variáveis que melhor explicam a diversidade de risco aceite e a forma como estas influenciam o risco seguro, definindo assim as classes tarifárias ajustadas e, se possível, o modelo que determina o prémio do contrato.

## 1.2. Conceitos Operatórios, terminologia

Antes de entrar na descrição dos dados sobre os quais foi desenvolvida esta tese, bem como na descrição da Teoria aplicada neste projeto, pensa-se ser da maior utilidade a apresentação de certos conceitos usados na atividade seguradora e que, de alguma forma, podem condicionar as opções de análise ou de desenvolvimento deste estudo.

No Seguro de Saúde existe um conjunto de coberturas possíveis: Internamento, Ambulatório, Parto, Estomatologia, Próteses e Ortóteses, Terapêuticas Não Convencionais, Medicamentos, Subsídio Diário, e outras de cariz mais acessório.

Todas estas coberturas têm uma componente de risco e uma componente de consumo, sendo que o Internamento será a que tem maior peso na primeira componente. Assim é muitas vezes definida como a cobertura base (obrigatória). Na Multicare muito poucos são os contratos que não têm Internamento. A segunda cobertura mais contratada é o Ambulatório e depois a Estomatologia, fazendo parte de uma cadeia de precedências estipuladas pela Multicare na comercialização dos seus seguros.

As principais coberturas deste tipo de Seguro são o Ambulatório, o Internamento e a Estomatologia, que, em conjunto, representam 90% dos custos com sinistros, no ramo Saúde.

O Internamento é a cobertura, que mais justifica a existência deste seguro, isto é, a mutualidade entre todos os clientes da Seguradora, já que os respetivos sinistros nesta cobertura têm uma probabilidade relativamente pequena – cerca de 3‰ – com uma grande variância nos respetivos custos, podendo atingir valores muito significativos (várias dezenas de milhar de euros).

No entanto, e talvez pela menor frequência de utilização, não é esta a cobertura que torna o seguro de saúde mais apelativo, é o **Ambulatório**. Esta tendência resulta da elevada frequência de utilização que o caracteriza e que, por abuso de linguagem, é referida habitualmente como "**cobertura com grande peso de consumo**", já que uma parte destes custos resulta de iniciativa própria e não de um acontecimento fortuito e alheio à vontade do próprio.

Ao valor total do custo do sinistro, independentemente da responsabilidade do respetivo pagamento ser do Cliente ou da Seguradora, chama-se **Valor Apresentado**. Neste trabalho, e porque se pretendem estudar as variáveis explicativas do custo da saúde, dentro do conjunto de características do seguro ou da pessoa que dele beneficia, optou-se por considerar como valor apresentado a soma de todos os custos apresentados no período de risco de cada indivíduo incluído na amostra. Este valor será designado por "**vap**"

A cada um dos indivíduos cobertos pelo seguro chama-se **Pessoa Segura**. Existe um titular que é o indivíduo pivô da relação e as restantes pessoas seguras que pertencem ao seu agregado familiar. O **titular** pode ser, ou não, o **Tomador do Seguro**, isto é o segundo contraente da apólice. A pessoa segura passa a ser **Cliente Utilizador** a partir do momento que usufruiu dos benefícios de alguma das coberturas.

Um dos indicadores utilizados na atividade seguradora é a **Taxa de Utilização**, que alguns autores também designam por **Taxa de Incidência,** visto ser um indicador com características semelhantes. Mas, de facto, pretende-se medir a utilização e, como tal, optou-se pela primeira designação. Esta frequência calcula-se dividindo o número de clientes utilizadores pelo número de Pessoas Seguras.

Existe um último conceito largamente utilizado que é o da **Anti-seleção**. Diz-se que há anti-seleção, sempre que a forma como o produto é apresentado ao público cria, de uma forma natural, maior interesse de aquisição aos clientes que aportam piores riscos. Por exemplo, os seguros Individuais produzem mais anti-seleção do que os seguros de grupo, porque estão naturalmente mais sensíveis à aquisição do seguro os agregados familiares onde existem pessoas menos saudáveis, enquanto nas grandes empresas o seguro de saúde é encarado como uma regalia .

## 1.3. Dados: Fontes e Limitações

Nos Seguros do Ramo Saúde, e, em especial, quando a utilização do Ambulatório se suporta de forma significativa em rede convencionada, não é claro o conceito de sinistro. Isto é, não se consegue identificar uma ocorrência, alheia à vontade da Pessoa Segura, que promova a realização de um conjunto de custos associados à reparação/indemnização dos danos causados por esse evento – o sinistro. Existem situações em que esses pagamentos estão dispersos por vários eventos do sistema operativo (ex: exames, consulta e tratamentos) e outras em que são registados num único evento (ex: tratamento de um dente).

Assim, e em vez de suportarmos a medição do risco em termos da frequência e do custo médio de sinistros, vamos fazê-lo através da probabilidade e do custo médio da utilização da cobertura:

$$Custo\ do\ Risco\ = Probabilidade\ de\ Utilização\ da\ Cobertura\ \times Custo\ Médio\ de\ Utilização$$

Os dados trabalhados foram os registos existentes na base de dados das Pessoas Seguras na Multicare no exercício de 2010 com exposição ao risco de um ano completo na cobertura de Ambulatório, expurgados de situações em que se levanta a suspeição de haver falha no preenchimento das variáveis consideradas passíveis de ter influência na determinação do risco.

Foram então selecionadas cerca de trezentas e oitenta e três mil (382.947) Pessoas Seguras na cobertura de Ambulatório, das quais:

- ✓ duzentas e quarenta e três mil (242.689) usufruíram da cobertura de Ambulatório,
- ✓ cento e quarenta mil (140.149) Pessoas Seguras não observaram qualquer sinistro
- ✓ as restantes cento e dez (110) surgem como tendo tido sinistros com valor apresentado nulo (todas as situações foram verificadas e são situações de erro de processamento – ocorrências de risco operatório – que foram posteriormente corrigidas e portanto são pessoas seguras que não utilizaram esta cobertura).

Os Seguros de Grupo têm habitualmente capitais seguros mais baixos. Apesar de se trabalhar com os valores apresentados, sabe-se que os clientes, a partir do momento em que têm conhecimento que esgotaram o plafond (valor do capital seguro disponível) deixam de registar despesas na Seguradora, ainda que as tenham. No entanto, como estamos sempre a

referir, do ponto de vista da Seguradora, não deixam de ser estes os valores apresentados à Companhia, embora não correspondam ao valor do risco.

# 2. Enquadramento teórico

A informação de cada pessoa segura, para além das características pessoais são ainda complementadas por toda a informação respeitante ao produto em vigor, bem como todos os dados referentes aos sinistros ocorridos no período em estudo.

Da panóplia de metodologias conhecidas para análise de dados distinguem-se os métodos de Regressão como processos de identificação de fatores explicativos da variável dependente. Assim, e tendo em consideração o objetivo deste estudo, a nossa abordagem teórica assentará essencialmente nos Métodos de Regressão Linear Generalizada (GLM).

Não obstante, serão utilizadas outras ferramentas, nomeadamente no que respeita à análise de dados multidimensionais.

## 2.1. Análise de *Clusters*

A Análise Classificatória é um método de agrupamento, quer de unidades estatísticas (indivíduos ou objetos), quer de variáveis em grupos, de tal forma que unidades situadas dentro do mesmo grupo são mais semelhantes do que unidades situadas em grupos distintos.

Trata-se de uma tipologia de análise exploratória que reduz a dimensão dos dados e cujo principal objetivo é classificar um conjunto de unidades estatísticas em grupos mutuamente exclusivos, exaustivos e homogéneos.

Será apresentado um resumo da terminologia respeitante a esta matéria que foi utilizada no desenvolvimento deste projeto.

### 2.1.1 Medidas de Dissemelhança

Para a construção de grupos de unidades estatísticas mais próximas, ou semelhantes, necessitamos de medidas de semelhança/dissemelhança ou distância.

A medida de **Semelhança/dissemelhança** é uma função que, a cada par de indivíduos faz corresponder o valor de um espaço euclidiano unidimensional (usualmente R). Habitualmente a semelhança e a dissemelhança são definidas em $[0,1] \subset R$ e linearmente opostas, sendo possível converter uma semelhança numa dissemelhança e vice-versa. Com uma medida de dissemelhança, a dois elementos semelhantes deve corresponder um valor baixo.

- **Propriedades da Dissemelhança**

**Seja** $d_{x_r.x_s.}$ a dissemelhança entre os individuos $x_r$ e $x_s$ respectivamente **então**

- $d_{x_r.x_s.} \geq 0, \ \forall \ x_r, x_s$
- $d_{x_r.x_r} = 0, \ \forall \ x_r$
- $d_{x_r.x_s.} = d_{x_s.x_r.}, \ \forall \ x_r, x_s$

**Se** a dissemelhança, para além das propriedades anteriores, ainda for marcada por:

- $d_{x_r.x_t.} + d_{x_t.x_s.} \geq d_{x_r.x_s.}, \ \forall \ x_r, x_s, x_t$
- $d_{x_r.x_s.} = 0$ sse $x_r = x_s$

**então** é ainda uma **distância** ou **métrica**.

- **Distância Euclidiana**

$$d_{x_r.x_s.} = \left[ \sum_{j=1}^{p} (x_{rj} - x_{sj})2 \right]^{1/2}$$

Esta medida calcula a distância entre dois indivíduos ($x_r$ e $x_s$ ) com base em cada uma das p variáveis e é uma das mais utilizadas por ser de fácil interpretação. Apresenta, no entanto, algumas **desvantagens**:

- ✓ Não é invariante às mudanças de escala, isto é, não deve ser utilizada quando as variáveis apresentam escalas muito distintas.
- ✓ Mostra um comportamento anómalo quando as variáveis apresentam variâncias muito distintas ou quando são correlacionadas.

- **Distância Euclidiana *Estandardizada***

Para ultrapassar estas dificuldades recorre-se muitas vezes à **Distância Euclidiana Estandardizada** ou **Distância de Karl Pearson**:

$d_{x_r.x_s.} = \left[ \sum_{j=1}^{p} \frac{(x_{rj} - x_{sj})^2}{s_j^2} \right]^{1/2}$, em que $s_j^2$ = desvio padrão da j-ésima variável

Ora esta não é mais do que a distância calculada sobre as variáveis estandardizadas.

### 2.1.2 Métodos de Classificação

Existem vários métodos de Classificação, sendo que os mais utilizados são os Hierárquicos ou os de Otimização.

Os **métodos Hierárquicos** podem ser Aglomerativos ou Divisivos: os primeiros partem das unidades estatísticas iniciais – variáveis ou indivíduos – para a agregação total e os segundos constroem-se no sentido inverso. Qualquer destes processos de associação, ou divisão, é suportado por uma regra de ligação, definida com base nas semelhanças/dissemelhanças anteriormente referidas e que permitem seleccionar a(s) classe(s) que se vão fundir, ou cindir, em cada passo.

- **Método da Ligação Simples ou do Vizinho mais próximo**

Este critério define como distância entre duas classes – $C_r$ e $C_s$- , a menor das distâncias entre quaisquer dois elementos pertencentes a esses grupos, ou seja:

$$D_{C_rC_s} = \text{mín}\,\{d_{rs}: r \in C_r \ \wedge s \in C_s\}$$

- **Método da Ligação Completa ou do Vizinho mais afastado**

Este critério define como distância entre duas classes – $C_r$ e $C_s$- , a maior das distâncias entre quaisquer dois elementos pertencentes a esses grupos, ou seja:

$$D_{C_rC_s} = \text{máx}\,\{d_{rs}: r \in C_r \ \wedge \ s \in C_s\}$$

- **Método da Ligação Média**

Este critério define como distância entre duas classes – $C_r$ e $C_s$- , a média das distâncias entre todos os pares de elementos pertencentes a esses grupos, ou seja:

$$D_{C_rC_s} = \frac{1}{n_r\, n_s}\sum_{r=1}^{n_r}\sum_{s=1}^{n_s} d_{rs} \ , \text{ em que : } \#\, C_r = n_r \wedge \#\, C_s = n_s$$

- **Método Centróide**

Este critério define como distância entre duas classes – $C_r$ e $C_s$- , a distância entre os respectivos centróides, ou seja:

$D_{C_rC_s} = d_{\bar{x}_r\bar{x}_s}$ , em que : $\bar{x}_r = \frac{1}{n_r}\sum_{r=1}^{n_r} x_r \wedge \bar{x}_s = \frac{1}{n_s}\sum_{s=1}^{n_s} x_s$ são os centróides

- **Método Ward**

Este critério baseia-se na perda de informação resultante do agrupamento dos indivíduos, perda que é medida através da soma dos quadrados dos desvios das observações individuais relativamente às médias dos grupos em que são classificadas.

Este método segue as seguintes fases:

- ✓ Cálculo das médias das variáveis para cada grupo;
- ✓ Cálculo do quadrado da distância Euclidiana entre essas médias e os valores das variáveis para cada indivíduo;
- ✓ Soma das distâncias para todos os indivíduos;
- ✓ Minimiza a variância dentro dos grupos. A função objetivo que se pretende minimizar é a soma dos quadrados dos erros.

- **Método da Ligação Mediana**

Este critério define como distância entre duas classes – $C_r$e $C_s$ – , a mediana das distâncias entre todos os pares de elementos pertencentes a esses grupos, ou seja:

$$D_{C_r C_s} = \text{mediana}\,\{d_{rs}: r \in C_r \ \wedge \ s \in C_s\}$$

Os **Métodos de Otimização** baseiam-se diretamente na escolha antecipada de um número de agrupamentos que conterão todos os casos. Procede-se, em seguida, a uma divisão de todos os casos pelos k grupos preestabelecidos e a melhor partição dos n casos será aquela que otimizar o critério escolhido. O Método Partitivo Iterativo usado para proceder a essa divisão, foi o denominado "*k-means*" ou "*nearest centroid sorting*". Segue os seguintes passos:

1. Começa por fazer uma partição inicial dos indivíduos por um número, predefinido pelo analista, de *Clusters*; calcula, para cada *cluster* o respetivo centróide;

2. Calcula as distâncias entre cada indivíduo e os centróides dos vários grupos; transfere cada indivíduo para o *cluster* relativamente ao qual se encontra a uma menor distância (por exemplo, distância Euclidiana);

3. Recalcula os centróides de cada *cluster*;

4. Repete os passos 2 e 3 até que todos os indivíduos se encontrem em *Clusters* estabilizados e que não seja possível efetuar mais transferências de indivíduos de um *cluster* para o outro.

### 2.1.3 Escolha do número de *Clusters*

Há situações em que se deve agrupar as unidades estatísticas até à classe única, mas para o objetivo pretendido neste trabalho ir-se-á selecionar o número de classes. Vários processos são conhecidos:

- Processo baseado na Distância Máxima

  Define-se um valor máximo admissível para a distância entre dois elementos incluídos no mesmo cluster e termina-se o processo quando qualquer elemento não enquadrado esteja a uma distância dos *clusters* superior ao limite estabelecido.

- Processo baseado no maior incremento

  Num método de classificação hierárquico aglomerativo, seja qual for a regra de fusão dos grupos ou a medida de dissemelhança que estamos a utilizar, os *clusters* vão-se construindo, associando os elementos mais próximos. Assim as distâncias entre eles vão crescendo. A regra determina que se pare a junção e se fique com o número de *clusters* existentes antes do maior dos incrementos, o que é visível no dendograma.

- **Processo baseado na variância explicada pela Classificação**

Seja X uma matriz constituída por n unidades estatísticas que se pretendem agrupar, segundo *p*, variáveis que a caracterizam:

$$X = [X_{ij}]_{\,i=1,\dots,n \,\wedge\, j=1,\dots,p}$$

Seja k o número de *Clusters* que se podem constituir com estes elementos, este processo consiste em:

*Passo 1 :* Determina-se, $\forall k = 1, \dots, n-1$, pelo método do *kmeans*, a classificação ótima das variáveis e, para cada uma delas, calcula-se a variabilidade explicada.

*Passo 2:* Assumindo que:

- ✓ $\{C_{r_1}, C_{r_2}, \dots, C_{r_k}\}$ é a classificação ótima encontrada para cada um dos valores de k,
- ✓ $\{n_{r_1}, n_{r_2}, \dots, n_{r_k}\}$ a dimensão de cada um deles
- ✓ $\{\bar{X}_{r_1}, \bar{X}_{r_2}, \dots, \bar{X}_{r_k}\}$ os respetivos centróides,

Teremos quatro tipos de variabilidade:

- × variabilidade dentro de cada cluster: $S_{r_l} = \sum_{j=1}^{p} \sum_{i=1}^{n_{r_l}} (X_{ij} - \bar{X}_{r_l})^2$
- × variabilidade dentro dos *clusters*[2]: $WSS_k = \sum_{l=1}^{k} S_{r_l}$
- × variabilidade Total da Amostra: $\text{TotSS(X)} = \sum_{j=1}^{p} \sum_{i=1}^{n} (X_{ij} - \bar{X})^2$ (1)
- × variabilidade explicada pela classificação:

$$BSS\{C_{r_1}, C_{r_2}, \dots, C_{r_k}\} = \text{TotSS(X)} - WSS_k \quad (2)$$

*Passo 3:* O número de *clusters* a selecionar será aquele que permite agrupar o máximo possível as unidades estatísticas, mantendo explicada uma elevada percentagem da variabilidade.

### 2.1.4 Avaliação da Classificação

[2] **Variabilidade perdida com a classificação.**

Depois de conseguida a classificação pode-se obter a medida da qualidade desta a partir de duas medidas, a saber:

- **Correlação Cofenética**

Este coeficiente, que se define pela expressão seguinte, indica uma boa classificação quando está próximo de 1:

$$r = \frac{\sum_i \sum_j (\delta_{ij} - \bar{\delta})(d_{ij} - \bar{d})}{n \times \sqrt{\sum_i \sum_j (d_{ij} - \bar{d})^2/(n-1)} \times \sqrt{\sum_i \sum_j (\delta_{ij} - \bar{\delta})^2/(n-1)}}$$

,onde $d_{ij}$ é a distância entre as unidades estatísticas iniciais, e

$\delta_{ij}$ é a distância entre as unidades, ou os grupos que integram, imediatamente antes da respetiva fusão

- **Medida da Variabilidade Explicada**

Esta medida assume que a integração de um elemento num cluster produz a sua substituição pelo centróide do grupo perdendo-se a variabilidade dessas unidades estatísticas face ao referido centróide. Assim a variabilidade explicada pela classificação será dada por: $\mathrm{BSS}\{C_{r_1}, C_{r_2}, \ldots, C_{r_k}\}/\mathrm{TotSS}(X)$ de acordo com as fórmulas (1) e (2) acima definidas.

## 2.2 Modelos de Regressão

Os modelos de Regressão podem ser utilizados para estudar a relação entre uma variável resposta (dependente) e um conjunto de variáveis explicativas (independentes) ou, de uma forma mais ambiciosa, pode ser utilizado como modelo de predição do valor esperado da variável.

### 2.2.1 Modelos Lineares Generalizados(GLM)

#### 2.2.1.1 Noções Introdutórias

A regressão linear tal como abordamos anteriormente faz parte de uma classe muito mais vasta de modelos que se designam Modelos Lineares Generalizados.

Denotamos na parte que se segue as variáveis resposta como $Y_1, Y_2, \ldots, Y_n$, as covariáveis do modelo como $Z_1, Z_2, \ldots, Z_q$ e o valor esperado da variável resposta $i$ como $\mu_i = E(Y_i \mid z_i)$, sendo $z_i = (1, z_{i1}, z_{i2}, \ldots, z_{iq})^T$ com $i = 1, \ldots, n$.

A extensão em relação ao modelo linear é feita em duas direções:

1) A distribuição considerada não tem de ser normal, podendo ser qualquer distribuição da família exponencial[3];

2) Em vez de existir uma relação linear direta entre $\mu_i$ e as covariáveis, nos GLM verificamos a relação de linearidade entre uma função diferenciável de $\mu_i$ – $g(\mu_i)$ – e as covariáveis, isto é, $g(\mu_i) = z_i^T \beta$.

As funções $g(\mu_i)$ e $\eta_i = z_i^T \beta$ são designadas como função de ligação e preditor linear, respetivamente.

A escolha da função de ligação depende do tipo de estudo que se quer fazer.

A função de ligação mais simples designa-se função de ligação canónica e tem a forma: $\theta_i = \eta_i = z_i^T \beta$, isto é, o parâmetro canónico coincide com o preditor linear.

Por exemplo, um caso particular dos GLMs amplamente conhecido é a regressão linear.

### 2.2.1.2 Modelos Lineares Generalizados – Distribuição Normal

Neste GLM temos que $Y_1, Y_2, ..., Y_n$ têm distribuição normal, assim a sua função de densidade pode escrever-se na forma:

$$f(y \mid \theta, \phi) = exp\left\{ \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \right\}.$$

Assim, se Y segue uma distribuição normal com valor médio μ e variância $\sigma^2$ a f.d.p. de Y é dada por:

$$f(y \mid \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(y-\mu)^2}{\sigma^2}} = exp\left\{ -ln\sqrt{\sigma^2 2\pi} - \frac{1}{2}\frac{(y-\mu)^2}{\sigma^2} \right\} = exp\left\{ -\frac{1}{\sigma^2}\left( \frac{y^2}{2} - \mu y + \frac{\mu^2}{2} \right) - ln\sqrt{\sigma^2 2\pi} \right\} =$$

$$= exp\left\{ -\frac{1}{\sigma^2}\left( -\mu y + \frac{\mu^2}{2} \right) - \frac{1}{\sigma^2} \times \frac{y^2}{2} - ln\sqrt{\sigma^2 2\pi} \right\} = exp\left\{ \frac{1}{\sigma^2}\left( \mu y - \frac{\mu^2}{2} \right) - \frac{1}{2}\left( \frac{y^2}{\sigma^2} + ln\left(\sqrt{\sigma^2 2\pi}\right)^2 \right) \right\} =$$

$$= exp\left\{ \frac{1}{\sigma^2}\left( \mu y - \frac{\mu^2}{2} \right) - \frac{1}{2}\left( \frac{y^2}{\sigma^2} + ln(\sigma^2 2\pi) \right) \right\}$$

Para $y \in \Re$. Fica-se então com:

---

[3] **Definição (Família Exponencial): Diz-se que uma variável aleatória Y tem distribuição pertencente à família exponencial se a sua função densidade de probabilidade ou função massa de probabilidade se puder escrever da forma:**

$$f(y \mid \theta, \phi) = exp\left\{ \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \right\}$$

**onde $\theta$ e $\phi$ são parâmetros escalares ($\theta$ de localização, $\phi$ de dispersão), sendo $\theta$ designado por parâmetro canónico. As funções $a(.)$, $b(.)$ e $c(.,.)$ são funções reais conhecidas em que $a(\phi) = \frac{\phi}{\omega}$ e $b(.)$ é diferenciável. Quando o suporte da distribuição não depende dos parâmetros estamos perante uma família regular. (Sem and Singer – 1993).**

**Para famílias regulares tem-se:** $E(Y) = \frac{\partial b(\theta)}{\partial \theta}$ **e** $Var(Y) = a(\phi) \times \frac{\partial^2 b(\theta)}{\partial \theta}$ **.**

$$\theta = \mu, \quad b(\theta) = \frac{\mu^2}{2}, \quad a(\phi) = \frac{\phi}{\varpi} = \sigma^2, \quad c(y,\phi) = -\frac{1}{2}\left(\frac{y^2}{\sigma^2} + ln\left(\sigma^2 2\pi\right)\right)$$

Tem-se assim como ligação canónica: $\theta_i = \eta_i = z_i^T \beta \Rightarrow \mu_i = z_i^T \beta$. Como resultado obtem-se a expressão tão familiar: $Y = X\beta + \varepsilon$

Uma das extensões mais naturais deste modelo é utilização da regressão linear múltipla com uma transformação da variável dependente – Y – permitindo assim alargar estes modelos a variáveis com distribuições com enviesamentos e caudas mais pesadas à direita.

## Modelos Lineares com Variável Endógena Transformada

Uma das transformações à variável endógena mais comum é a logarítmica que permite estender a aplicabilidade destes modelos a variáveis com distribuição Lognormal.

Estes modelos, apesar de serem conceptualmente simples, trazem algumas complicações de cariz prático relacionado essencialmente com a interpretação e com a construção de intervalos de confiança.

Por exemplo, quando se analisam custos não é admissível apresentar resultados em Log de euros, mas também não se pode simplesmente aplicar a função inversa, já que $E[Y|X] \neq \exp\{E[Ln(Y)|X]\}$. Em alternativa, para uma transformação logarítmica e assumindo a normalidade dos desvios, o valor esperado do custo será dado por:

- $E[Y|X] = \exp(X^T\beta + 0{,}5\sigma_\varepsilon^2) = \exp(X^T\beta) \times \exp(0{,}5\sigma_\varepsilon^2)$

Onde X representa o vector das variáveis explicativas, β o vector com p+1 coeficientes de regressão (j=0,…,p) e $\varepsilon$ é o erro aleatório[4].

Quando a distribuição do erro não é Normal, mas estes são homocedásticos, pode-se utilizar o estimador "*smearing*" de Duan (Duan,1983) para a transformação Logarítmica:

- $E[Y|X] = \varphi \times \exp(X^T\beta)$ , onde o fator *smearing* é dado por:
  $\hat{\varphi} = n^{-1}\sum_{i=1}^{n}\exp(\hat{\varepsilon}_i)$, em que $\hat{\varepsilon}_i = \ln Y_i - X_i^T\hat{\beta}$

Em aplicações práticas é altamente improvável que seja sustentável para os dados de custos individuais a suposição de os erros serem homocedásticos. Se os erros são heteroscedásticos então estimador "*smearing*" de Duan será tendencioso. No entanto, se a forma de heterocedasticidade for uma função p(x) (dos regressores X) conhecida, então as previsões imparciais de custo é dada como:

$$E[Y|X] = p(X) \times \exp(X^T\beta)$$

Quando a variância é função dos vários regressores e estes são contínuos e não discretos, a especificação da forma da heterocedasticidade pode ser problemática, já que a forma exata é muitas vezes desconhecida. Nesses casos, pode ser útil calcular o estimador "smearing" para

[4] **Nota de publicação de Santos Silva e Tenreyo, 2006: No caso de os erros serem heterocedásticos o método dos mínimos quadrados pode produzir estimativas enviesadas para os coeficientes**

percentis da gamma ajustada a partir do modelo[5] e de seguida, dividir a distribuição dos valores ajustados em cinco intervalos, usando os percentis, e calcular os fatores "*smearing*" resultantes.

Para a transformação da Raiz quadrada dos custos podemos utilizar uma correção à transformação inversa semelhante. Neste caso, contudo, o termo de correção deverá ser aditivo no caso da homocedasticidade: $E[Y|X] = (T^T\beta)^2 + \varphi$

, onde o fator smearing é dado por: $\widehat{\varphi} = n^{-1}\sum_{i=1}^{n} \hat{\varepsilon}_i^{\,2}$ , e: $\hat{\varepsilon}_i = \sqrt{Y_i} - X_i^T\hat{\beta}$

Para o estudo que se pretende, vamos estudar ainda outros dois tipos de GLM: um modelo para variável resposta binária – a utilização ou não do seguro – e um modelo para variável resposta continua – o custo da utilização quando esta existir.

### 2.2.1.3 Modelos Lineares Generalizados – Modelo Logístico e Modelo Gama

#### Regressão Logística

O modelo de Regressão Logística é adequado para dados binomiais e em particular, para dados *Bernoulli*.

- *Dados Binomiais*

Se Y segue uma distribuição binomial com parâmetros $m$ e $\pi$ ($Y \sim Bin(m,\pi)$), a sua f.m.p. é dada por:

$$f(y\mid\pi) = \binom{m}{y}\pi^y(1-\pi)^{m-y} = exp\left\{y\,ln\pi + (m-y)ln(1-\pi) + ln\binom{m}{y}\right\} =$$

$$= exp\left\{y\left(ln\pi - ln(1-\pi)\right) + mln(1-\pi) + ln\binom{m}{y}\right\} = exp\left\{y\,ln\left(\frac{\pi}{1-\pi}\right) + mln(1-\pi) + ln\binom{m}{y}\right\}$$

Diz-se assim que a variável aleatória Y tem distribuição pertencente à família exponencial com:

$$\theta = ln\left(\frac{\pi}{1-\pi}\right),\quad b(\theta) = -m\,ln(1-\pi) = m\,ln(1+e^{\theta}),\quad c(y,\phi) = ln\binom{m}{y},\quad a(\phi) = 1 \Rightarrow \phi = \varpi = 1$$

Especificamente no caso da *Bernoulli*, isto é Binomial(1,p), temos :

- *Dados Bernoulli*

$$f(y|\pi) = \exp\left\{y\ln\left(\frac{\pi}{1-\pi}\right) + \ln(1-\pi)\right\}$$ , com:

[5] **Utilizar os percentis de exp($X^T\hat{\beta}$)**

$$\theta = ln\left(\frac{\pi}{1-\pi}\right), \quad b(\theta) = -\ln(1-\pi) = \ln(1+e^{\theta}), \quad c(y,\phi) = 0, \quad a(\phi) = 1 \Rightarrow \phi = \varpi = 1$$

Assim, considerando as variáveis resposta $Y_i \sim Be(\pi_i)$, com $E(Y_i) = \pi_i$ e $\theta_i = ln\left(\frac{\pi_i}{1-\pi_i}\right)$, temos como ligação canónica (logit):

$$\theta_i = \eta_i = z_i{}^T \beta \Rightarrow ln\left(\frac{\pi_i}{1-\pi_i}\right) = z_i{}^T \beta$$

E para a probabilidade de sucesso, $P(Y_i = 1) = \pi_i$, temos a relação: $\pi_i = \frac{\exp(z_i{}^T \beta)}{1+\exp(z_i{}^T \beta)}$

A função $F(x) = \frac{\exp(x)}{1+\exp(x)}$, tal que $F: \Re \to [0,1]$, é a função de distribuição logística. Por esse motivo, o GLM definido pelo modelo binomial, e em particular o *Bernoulli*, com função de ligação canónica é conhecido por modelo de regressão logística.

## Regressão Gamma

O modelo de Regressão Gama é usado na análise de dados contínuos com suporte positivo para a variável resposta. Estes modelos adaptam-se quando estamos perante um coeficiente de variação[6] constante.

- *Dados Gamma*

Se Y segue uma distribuição Gama com parâmetros $\upsilon$ e $\mu$ ($Y \sim Gamma(\upsilon/\mu, \upsilon)$), a sua f.m.p. é dada por:

[6] Amaral Turkman, M.A. e Silva, G. (2000) O coeficiente de variação define-se como : CV(X) = $\sqrt{Var[X]}/E[X]$

$$f(y|\pi) = \exp\left\{\upsilon\, ln\left(\frac{\nu}{\mu}\right) + (\nu-1)\ln(y) - \frac{\upsilon}{\mu}\, ln(y) - \ln\left(\Gamma(\nu)\right)\right\}$$
$$= exp\left\{\upsilon\, ln\left(\nu\right) - \nu\ln(\mu) + (\nu-1)\ln(y) - \frac{\upsilon}{\mu}\, y - \ln\left(\Gamma(\nu)\right)\right\}$$
$$= exp\left\{\nu \times \left({}^{-1}/_{\mu}\right)y - \upsilon\ln(\mu) + \nu\ln(\nu) + (\nu-1)\ln(y) - \ln\left(\Gamma(\nu)\right)\right\}$$
$$= exp\left\{\frac{\left({}^{-1}/_{\mu}\right)y - \ln(\mu)}{{}^{1}/_{\nu}} + \nu\ln(\nu) + (\nu-1)\ln(y) - \ln\left(\Gamma(\nu)\right)\right\}$$

, com:

$$\theta = {}^{-1}/_{\mu}, \quad b(\theta) = \ln(\mu) = -\ln(-\theta), \quad c(y,\phi) = \nu\ln(\nu) + (\upsilon-1)\ln(y) - \ln(\Gamma(\upsilon)),$$
$$a(\phi) = 1/\upsilon \Rightarrow \phi = 1/\upsilon \wedge \varpi = 1$$

Assim, considerando as variáveis resposta $Y_i \sim Gamma({}^{\upsilon}/_{\mu_i}, \upsilon)$, com $E(Y_i) = \mu_i$ e admitindo $\mu_i = \exp\left\{z_i{}^T\beta\right\}$, obtemos um modelo linear generalizado gama, dado que:

- as varáveis resposta são independentes,
- a distribuição é da família exponencial, com $\theta_i = \frac{-1}{\exp(z_i{}^T\beta)}$, $\phi = \frac{1}{\nu}$ e $\varpi_i = 1$
- o valor esperado μi está relacionado com o predictor linear ηi através da relação μi = exp(ηi),
- a função ligação é a função logarítmica.

E o modelo poderá ser escrito da forma:

$$\boldsymbol{Y_i = \exp(z_i{}^T\beta)\epsilon_i,\ i=1,\ldots,n \wedge \epsilon_i \text{ i.i.d. } Gama(\nu,\nu)}$$

### 2.2.2 Estimação dos Parâmetros

O método de estimação mais comum é o método de máxima verosimilhança. Sendo assim, temos como função de verosimilhança:

**Regressão Logística**

$$L(\beta) = \prod_{i=1}^{n} f(y_i|\pi_i) = exp\left\{\sum_{i=1}^{n} y_i \ln(\frac{\pi_i}{1-\pi_i}) - \sum_{i=1}^{n}\ln(1-\pi_i)\right\}$$

Escrevendo a expressão anterior com $\pi_i = \frac{\exp(z_i^T\beta)}{1+\exp(z_i^T\beta)}$ ficamos com:

$$L(\beta) = exp\left\{\sum_{i=1}^{n} y_i \ln\left(\frac{\frac{\exp(z_i^T\beta)}{1+\exp(z_i^T\beta)}}{1-\frac{\exp(z_i^T\beta)}{1+\exp(z_i^T\beta)}}\right) - \sum_{i=1}^{n}\ln\left(1-\frac{\exp(z_i^T\beta)}{1+\exp(z_i^T\beta)}\right)\right\} =$$

$$exp\left\{\sum_{i=1}^{n} y_i \ln\left(\frac{\frac{\exp(z_i^T\beta)}{1+\exp(z_i^T\beta)}}{\frac{1+\exp(z_i^T\beta)-\exp(z_i^T\beta)}{1+\exp(z_i^T\beta)}}\right) - \sum_{i=1}^{n}\ln\left(\frac{1+\exp(z_i^T\beta)-\exp(z_i^T\beta)}{1+\exp(z_i^T\beta)}\right)\right\} =$$

$$exp\left\{\sum_{i=1}^{n} y_i \ln\left(\frac{\frac{\exp(z_i^T\beta)}{1+\exp(z_i^T\beta)}}{\frac{1}{1+\exp(z_i^T\beta)}}\right) - \sum_{i=1}^{n}\ln\left(\frac{1}{1+\exp(z_i^T\beta)}\right)\right\} =$$

$$exp\left\{\sum_{i=1}^{n} y_i \ln\left(\frac{\exp(z_i^T\beta)}{1}\right) - \sum_{i=1}^{n}\ln(1+\exp(z_i^T\beta))\right\} = exp\left\{\sum_{i=1}^{n} y_i \times z_i^T\beta - \sum_{i=1}^{n}\ln(1+\exp(z_i^T\beta))\right\}$$

Aplicando o logaritmo à função de verosimilhança (que chamamos log-verosimilhança):

$$l(\beta) = \ln[L(\beta)] = \sum_{i=1}^{n} y_i \sum_{j=0}^{p} z_{ij}\beta_j - \sum_{i=1}^{n}\ln(1+e^{\sum_{j=0}^{p} z_{ij}\beta_j})$$

Derivando a função log-verosimilhança, temos que os estimadores de máxima verosimilhança para $\beta$ são obtidos como solução do sistema de equações:

$$\frac{\partial \ln L(\beta)}{\partial \beta_j} = \sum_{i=1}^{n}\left\{y_i z_{ij} - \frac{\exp(z_i^T\beta)}{1+\exp(z_i^T\beta)} z_{ij}\right\} = 0, \quad j = 0,\ldots,p$$

Uma vez que não é possível encontrar a solução do sistema analiticamente, é necessário recorrer a métodos numéricos.

## Regressão Gamma

Partindo da expressão do modelo Gama para a construção dos estimadores de máxima verosimilhança:

$$L(\beta) = \prod_{i=1}^{n} f(y_i|\pi_i) = exp\left\{\frac{(-1/\mu_i)y - \ln(\mu_i)}{1/\nu} + \nu\ln(\nu) + (\nu-1)\ln(y) - \ln(\Gamma(\nu))\right\}$$

Escrevendo a expressão anterior com $\mu_i = \exp(z_i^T \beta)$ ficamos com:

$$L(\beta) = exp\left\{\sum_{i=1}^{n}\left[\frac{-y_i/\exp(z_i^T\beta) - \ln\left(\exp(z_i^T\beta)\right)}{1/\nu} + \nu\ln(\nu) + (\nu-1)\ln(y) - \ln\left(\Gamma(\nu)\right)\right]\right\}$$

$$= exp\left\{\sum_{i=1}^{n}\left[\frac{-y_i/\exp(z_i^T\beta) - z_i^T\beta}{1/\nu}\right] + n\nu\ln(\nu) + n(\nu-1)\ln(y) - \mathrm{n}\times\ln\left(\Gamma(\nu)\right)\right\}$$

Aplicando o logaritmo à função de verosimilhança (que chamamos log-verosimilhança):

$$l(\beta) = \ln[L(\beta)] = \sum_{i=1}^{n}\left[\frac{-y_i/\exp(\sum_{j=0}^{p} z_{ij}^T\beta_j) - \sum_{j=0}^{p} z_{ij}^T\beta_j}{1/\nu}\right] + n\nu\ln(\nu) + + n(\nu-1)\ln(y) - \mathrm{n}\times\ln\left(\Gamma(\nu)\right)$$

Derivando a função log-verosimilhança, temos que os estimadores de máxima verosimilhança para $\beta$ são obtidos como solução do sistema de equações:

$$\frac{\partial l(\beta)}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial}{\partial \beta_j}\left[\frac{-y_i/\exp(\sum_{j'=0}^{p} z_{ij'}{}^T\beta_{j'}) - \sum_{j'=0}^{p} z_{ij}{}^T\beta_{j'}}{1/\nu}\right] =$$

$$= \nu \sum_{i=1}^{n}\left[-y_i \frac{\partial}{\partial \beta_j}\exp\left(\sum_{j'=0}^{p} z_{ij'}{}^T\beta_{j'}\right)^{-1} - z_{ij}\right] =$$

$$= \nu \sum_{i=1}^{n}\left[y_i \exp\left(\sum_{j'=0}^{p} z_{ij'}{}^T\beta_{j'}\right)^{-2} \frac{\partial}{\partial \beta_j}\exp\left(\sum_{j'=0}^{p} z_{ij'}{}^T\beta_{j'}\right) - z_{ij}\right] =$$

$$= \nu \sum_{i=1}^{n}\left[y_i \exp\left(\sum_{j'=0}^{p} z_{ij'}{}^T\beta_{j'}\right)^{-2} \exp\left(\sum_{j'=0}^{p} z_{ij'}{}^T\beta_{j'}\right) z_{ij} - z_{ij}\right] =$$

$$= \nu \sum_{i=1}^{n}\left[y_i \exp\left(\sum_{j'=0}^{p} z_{ij'}{}^T\beta_{j'}\right)^{-1} z_{ij} - z_{ij}\right] =$$

$$= \nu \sum_{i=1}^{n}\left[z_{ij}\left[y_i \exp\left(\sum_{j'=0}^{p} z_{ij'}{}^T\beta_{j'}\right)^{-1} - 1\right]\right] = 0$$

Mais uma vez se chega a uma expressão para a qual não é possível encontrar a solução do sistema analiticamente, assim temos de recorrer a métodos iterativos[7].

Existem vários métodos destes, que podem ser utilizados para encontrar o maximizante de $f(\beta) = \ln L(\beta)$. Os dois mais referenciados são os que se seguem:

- ***Método de Newton-Raphson***

Quando temos n equações não lineares, baseia-se no desenvolvimento em série de *Taylor*, $f(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!}(x-a)^n$, até à segunda ordem com $a = x_0$ :

- Com n=1: $f(x) = f(x_0) + f'(x_0)(x-x_0) + f''(x_0)\frac{(x-x_0)^2}{2!} + o\left[(x-x_0)^3\right]$.

O máximo da função $f(x)$ deve estar próximo do máximo de $f^*(x) = f(x_0) + f'(x_0)(x-x_0) + f''(x_0)\frac{(x-x_0)^2}{2!}$. Assim, e como a log-verosimilhança de um

[7] O Software utilizado para realizar as estimações foi o R e este utiliza o Método Iterativo dos Mínimos Quadrados.

modelo Logit é uma função côncava[8], basta-nos encontrar a solução de $\frac{\partial f^*(x)}{dx} = 0$:

$$f'(x_0)+(x-x_0)f''(x_0)=0 \Leftrightarrow x = x_0 - \frac{f'(x_0)}{f''(x_0)}$$

Este valor $x$ é agora usado para melhorar a aproximação. Chegamos assim ao método iterativo em que se repete o processo:

$$\beta_{k+1} = \beta_k - \frac{f'(\beta_k)}{f''(\beta_k)}$$

- Com n>1:

Neste caso, o método tem a forma: $\beta_{k+1} = \beta_k - H(\beta_k)^{-1} J(\beta_k)$, em que $H(\beta_k)$ é a matriz Hessiana de f, $J(\beta_k)$ é a matriz Jacobiana de f, e $\beta_k$ o vector de parâmetros estimado na k-ésima iteração. Isto é, $\{H(\beta)\}_{ij} = \frac{\partial^2 f(\beta)}{d\beta_i d\beta_j}$ e $\{J(\beta)\}_j = \frac{\partial f(\beta)}{d\beta_j}$.

As desvantagens deste método passam, essencialmente, pela necessidade do cálculo e inversão da matriz Hessiana em cada iteração e pela necessidade de boas estimativas iniciais. Caso contrário, na maioria dos casos não há garantia da convergência do método para o máximo global.

[8] Amaral Turkman, M.A. e Silva, G. (2000), pág:46 a 48 - Todos os modelos utilizados neste estudo apresentam estimadores de máxima verosimilhança finitos, com estimativas no interior do espaço paramétrico e únicos.

- ***Método Iterativo dos Mínimos Quadrados (Fisher's Scoring method)***

Este método pode ser considerado uma variante estatística do método *Newton-Raphson*. A grande diferença consiste na substituição da segunda derivada (n=1), ou matriz Hessiana (n>1) pelo seu valor esperado.

De acordo com isto, algumas definições importantes são:

- Função score: $S(\beta)=\frac{\partial \ln L(\beta)}{\partial \beta}$.

Para famílias regulares temos que:

$$E\left(S(\beta)\right)=0 \quad e \quad E\left(S^T(\beta)S(\beta)\right)=-E\left[\frac{\partial^2 \ln L(\beta)}{\partial\beta\,\partial\beta^T}\right]$$

- Matriz de Informação de Fisher: $I(\beta)=E\left[-\frac{\partial S(\beta)}{\partial\beta}\right]=E\left[-\frac{\partial^2 \ln L(\beta)}{\partial\beta\ \partial\beta^T}\right]$

A matriz de informação de Fisher coincide com o simétrico da matriz Hessiana.

Chegamos assim ao método iterativo em que se repete o processo:
$\beta_{k+1}=\beta_k+I(\beta_k)^{-1}S(\beta_k)$, em que os valores de $\beta_k$ são as estimativas de $\beta$ na k-ésima iteração.

Um critério de paragem comum para os dois processos é limitar o erro absoluto, ou seja, definir um valor para ε tal que quando se obtém $\|x_k-x_{k-1}\|<\varepsilon$, o método é interrompido e considera-se como solução $x_k$.

### 2.2.3 Propriedades dos Parâmetros Estimados

Para fazermos inferências sobre os parâmetros estimados, é necessário conhecer a distribuição de $\hat{\beta}$. Uma vez que não é possível, em geral, obter as distribuições de amostragem exactas para os estimadores, utilizam-se resultados assintóticos.

Com as propriedades já mencionadas do vector score $S(\beta)$, pelo Teorema do Limite Central, temos $S(\beta)\xrightarrow{L}N_{q+1}(0,I(\beta))$ e consequentemente, $S(\beta)^T I^{-1}(\beta)S(\beta)\xrightarrow{L}\chi^2_{q+1}$.

Desenvolvendo $S(\beta)$ em Série de *Taylor* até à 1ª ordem, em torno de $\hat{\beta}$ obtemos:

$$S(\beta)\approx S(\hat{\beta})+\frac{\partial S(\beta)}{d\beta}|_{\beta=\hat{\beta}}(\beta-\hat{\beta})$$

Temos que $S(\hat{\beta})=0$ e $\frac{\partial S(\beta)}{d\beta}|_{\beta=\hat{\beta}}=H(\hat{\beta})$ e considerando $H(\hat{\beta})=-I(\beta)$ ficamos com:

$$S(\beta) \approx -I(\beta)(\beta - \hat{\beta}) \Rightarrow \hat{\beta} - \beta \approx I^{-1}(\beta)S(\beta)$$

Com a expressão anterior é agora possível deduzir as propriedades assintóticas dos estimadores de máxima verosimilhança de $\beta$:

- $E(\hat{\beta}) \approx \beta$, $\hat{\beta}$ é um estimador aproximadamente centrado de $\beta$;
- $\text{cov}(\hat{\beta}) \approx E\left[\left(\hat{\beta} - \beta\right)\left(\hat{\beta} - \beta\right)^T\right] = I^{-1}(\beta)$
- $\hat{\beta} \xrightarrow{L} N_p(\beta, I^{-1}(\beta))$
- A estatística de Wald $(\hat{\beta} - \beta)^T I(\beta)(\hat{\beta} - \beta) \xrightarrow{L} \chi^2_{q+1}$

Temos ainda que para o modelo em estudo, a estimativa de máxima verosimilhança de $\beta$ existe no interior do espaço paramétrico, é finita e única. [8]

### 2.2.4 Validação dos Modelos

- ***Significância do modelo***

Para testar se as variáveis independentes são significativamente explicativas, é necessário realizar testes sobre o parâmetro $\beta$, que podem ser formulados na forma:

$H_0 : C\beta = \xi \quad vs \quad H_1 : C\beta \neq \xi$

onde C é uma matriz $p \times (q+1)$, com $p \leq q+1$ de característica $p$.

Existem essencialmente duas estatísticas para testar as hipóteses deste tipo:

- Estatística de Wald:

Como já foi abordado $\hat{\beta} \xrightarrow{L} N_{q+1}(\beta, I^{-1}(\beta))$. Consequentemente, uma vez que $C\hat{\beta}$ é uma transformação linear de $\hat{\beta}$, temos $C\hat{\beta} \xrightarrow{L} N_p(C\beta, C\,I^{-1}(\beta)C^T)$.

Estatística de Teste: sob $H_0$ temos:

$W = (C\hat{\beta} - \zeta)^T \left[C\,I^{-1}(\hat{\beta})C^T\right]^{-1}(C\hat{\beta} - \zeta) \xrightarrow{L} \chi^2_p$.

Região de Rejeição: $\left\{W_{obs} > \chi^{1-\alpha}_p\right\}$.

Utilidade: Principalmente testar hipóteses nulas sobre componentes individuais. Nestes casos ficamos com:

$H_0 : \beta_j = 0$ vs $H_1 : \beta_j \neq 0$

Estatística de Teste: sob $H_0$ temos:

$W = \dfrac{\hat{\beta}^2}{\sigma_{jj}} \xrightarrow{L} \chi^2_1$, sendo $\sigma_{jj}$ o j-ésimo elemento da diagonal de $I^{-1}(\hat{\beta})$

Região de Rejeição: $\left\{W_{obs} > \chi^{1-\alpha}_1\right\}$.

- Estatística de Wilks ou de Razão de Verosimilhanças:

Consideremos $\tilde{\beta}$ como o estimador de máxima verosimilhança restrito, isto é, como o valor de $\beta$ que maximiza a verosimilhança sujeita a $H_0 : C\beta = \xi$ .

Estatística de Teste: sob $H_0$ temos[9]:

$$K = -2\ln\frac{max_{H_0} L(\beta)}{max_{H_0 \cup H_1} L(\beta)} = -2\left\{\ln L(\tilde{\beta}) - \ln L(\hat{\beta})\right\} \xrightarrow{L} \chi^2_p$$ .10

Região de Rejeição: $\left\{K_{obs} > \chi_p^{1-\alpha}\right\}$.

Utilidade: Comparar modelos encaixados.

- ***Qualidade do modelo***

  - Análise dos Resíduos

Tal como na Análise de Regressão Linear, para avaliar a qualidade do modelo é importante analisarmos os resíduos.

Os resíduos de Pearson são dados por:

$$r_i = \frac{(y_i - \hat{\mu}_i)\varpi_i}{\sqrt{\hat{\phi} V(\hat{\mu}_i)}}$$ , i=1,...,n

**Para o modelo Logístico** com $Y \sim Ber(\pi)$ temos $\hat{\mu}_i = \pi_i$, $V(\hat{\mu}_i) = \pi_i(1-\pi_i)$, $\phi = \varpi = 1$ ficamos com:

$$r_i = \frac{(y_i - \pi_i)}{\sqrt{\pi_i(1-\pi_i)}}$$

E no caso particular **do modelo Gama** com $Y_i \sim Gama(\upsilon, \upsilon/\mu_i)$ temos, $V(\hat{\mu}_i) = \mu_i^2$, $\phi = 1/\upsilon$, $\varpi = 1$ ficamos com:

$$r_i = \frac{(y_i - \mu_i)}{\mu_i/\sqrt{\upsilon}}$$

Os resíduos de Pearson padronizados são dados por:

$$r_i^P = \frac{X_i}{\sqrt{1-h_i}}$$,

com $h_i$ sendo o elemento i da diagonal da matriz "hat", tal como está definida no Anexo

A desvantagem da utilização dos resíduos de Pearson é que a sua distribuição é, geralmente, muito assimétrica para modelos não normais.

Com base nos resíduos e verosimilhanças, é possível ainda avaliar a qualidade do modelo e comparar vários modelos com base em algumas medidas como:

  - Função Desvio- *Deviance* $-D(y;\hat{\mu})$

[9] Pelo Teorema de Wilks (Cox and Hinkley, 1974)

[10] O número de graus de liberdade corresponde à diferença entre o número de parâmetros a estimar sob $H_0 \cup H_1$ (neste caso q+1) e o número de parâmetros a estimar sob $H_0$ (neste caso q+1-p).

Esta medida é baseada na estatística de razão de Verosimilhanças, avalia a discrepância entre o modelo saturado – S (modelo com tantos parâmetros quanto observações) e o modelo corrente – M.

Obtemos assim a estatística:

$$D^*(y;\hat{\mu}) = -2\left\{\ln L_M(\tilde{\beta}) - \ln L_S(\hat{\beta})\right\} = \frac{D(y;\hat{\mu})}{\phi}$$

**Considerando o modelo Logístico** (no caso particular da *Bernoulli*) temos que para o modelo saturado cada parâmetro $\pi_i$ é estimado com base no valor real observado, isto é, $\hat{\pi}_i = y_i$. Para o modelo corrente, temos q+1 parâmetros, $q+1<n$, e os valores de $\hat{\pi}_i$ são estimados com recurso aos valores ajustados, isto é, $\hat{\pi}_i = \hat{y}_i$. Fazendo estas substituições e considerando $\phi = 1$ ficamos com:

$$\begin{aligned}
D(y;\hat{\mu}) &= -2\left\{\ln L_M(\tilde{\beta}) - \ln L_S(\hat{\beta})\right\} = \\
&= -2\sum_{i=1}^{n}\left\{\left[y_i \ln\left(\frac{\hat{y}_i}{1-\hat{y}_i}\right) + \ln(1-\hat{y}_i)\right] - \left[y_i \ln\left(\frac{y_i}{1-y_i}\right) + \ln(1-y_i)\right]\right\} = \\
&= -2\sum_{i=1}^{n}\{\{y_i \ln(\hat{y}_i) - y_i \ln(1-\hat{y}_i) + \ln(1-\hat{y}_i)\} - \{y_i \ln(y_i) - y_i \ln(1-y_i) + \ln(1-y_i)\}\} = \\
&= -2\sum_{i=1}^{n}\{\{y_i \ln(\hat{y}_i) + (1-y_i)\ln(1-\hat{y}_i)\} - \{y_i \ln(y_i) + (1-y_i)\ln(1-y_i)\}\} \\
&= -2\sum_{i=1}^{n}\left\{y_i \ln\left(\frac{\hat{y}_i}{y_i}\right) + (1-y_i)\ln\left(\frac{1-\hat{y}_i}{1-y_i}\right)\right\}
\end{aligned}$$

De novo, e **considerando agora o modelo Gama**, o modelo saturado S com $\hat{\mu}_i = y_i$ e o modelo corrente M (de q+1 parâmetros, $q+1<n$) com $\hat{\mu}_i = \hat{y}_i$. Ficamos com:

$$\begin{aligned}
D(y;\hat{\mu}) &= -2\left\{\ln L_M(\tilde{\beta}) - \ln L_S(\hat{\beta})\right\} = \ldots = \\
&= -2\sum_{i=1}^{n}\left\{\left(\frac{y_i - \hat{y}_i}{\hat{y}_i}\right) - \ln\left(\frac{y_i}{\hat{y}_i}\right)\right\} \quad \text{, e } \phi = 1/\upsilon
\end{aligned}$$

Considerando modelos com o mesmo número de parâmetros, o melhor modelo é o que apresenta um menor desvio[11] .

Uma das grandes vantagens da função desvio, quando temos de trabalhar com uma combinação de vários modelos e não com um único modelo, é a sua aditividade como medida de discrepância, apesar de não ter uma interpretação directa como tem a estatística de Pearson.

- Critério de informação de Akaike (AIC)

Este critério, tal como a Função Desvio, é baseado na função de log-verosimilhança. No entanto, uma vez que nos interessa não só um modelo que se ajuste bem, mas um modelo

[11] De forma análoga pode também ser considerada como a discrepância entre o modelo corrente – M e o modelo Nulo – N. Neste caso, o melhor modelo é o que apresenta maior desvio.

parcimonioso, esta medida tem um factor de penalização para o número de parâmetros do modelo.

A estatística vem dada por:

$$AIC = -2l(\tilde{\beta}_1) + 2\,dim(\tilde{\beta}_1)$$

O ajustamento do modelo é tanto melhor quanto menor for o AIC.

- Curva ROC (para aplicação em modelos binários)

A curva ROC é uma gráfico da probabilidade de se detectar os verdadeiros positivos(sensibilidade) e os verdadeiros negativos (1-especificidade) para diferentes pontos de corte.

Comecemos por definir sensibilidade e especificidade. A sensibilidade é a capacidade do modelo estimar um resultado positivo – no nosso caso estimar uma utilização do seguro – quando o indivíduo de facto o utilizou.

A especificidade é, em simetria, a capacidade do modelo estimar uma não utilização do seguro para um indivíduo que de facto não o utilizou.

Definamos agora a seguinte tabela:

| | U+ | U- | |
|---|---|---|---|
| GLM+ | a | b | VPP |
| GLM- | c | d | VPN |
| | **S** | **E** | |

**Tabela 1:** Classificação dos indivíduos em função das concordâncias entre o observado e a estimativa do modelo

, em que:

"a" será o número de indivíduos que utilizaram o seguro [U+] e o modelo determina uma utilização [GLM+],

"b" será o número de indivíduos que não utilizaram o seguro [U-] mas o modelo determina uma utilização [GLM+],

"c" será o número de indivíduos que utilizaram o seguro [U+] mas o modelo determina uma não utilização [GLM-],

"d" será o número de indivíduos que não utilizaram o seguro [U-] e o modelo determina uma utilização [GLM+-],

Assim:

S = a/(a + c) define a sensibilidade do modelo,

E = d/(b + d) define a especificidade do modelo,

VPP = a/(a + b) define o Valor Preditivo Positivo do modelo, isto é a probabilidade de o indivíduo utilizar o seguro dado que o modelo o prevê, e

VPN = d/(c + d) define o Valor Preditivo Negativo do modelo, isto é a probabilidade de o indivíduo não utilizar o seguro dado que o modelo prevê uma não utilização.

De notar que, como o modelo logístico determina a probabilidade de utilização do seguro, assim temos de definir o "cut-off" de predição da utilização, isto é, qual é a probabilidade de utilização que corresponde a uma utilização.

Desta forma a curva ROC, vai corresponder ao par (1-especificidade , sensibilidade) para todos os valores possíveis do cut-off. A área sob esta curva (AUC) resultará num valor em [0,1] determinando a qualidade do modelo da seguinte forma:

| AUC | Diagnóstico |
|---|---|
| **= 0,5** | Modelo sem poder discriminatório |
| **0,7 ≤ AUC ≤ 0,8** | Discriminação aceitável |
| **0,8 ≤ AUC ≤ 0,9** | Discriminação excelente |
| **AUC ≥ 0,9** | Discriminação extraordinária |

**Tabela 2:**Diagnósticos em função do AUC

- Teste de Park

Na confirmação da distribuição associada ao Modelo GLM o "Teste de Park"(Park,1966) tem se demonstrado uma ferramenta útil em aplicações empíricas:

Este teste consiste em fazer uma regressão do quadrado dos resíduos (do MLG ou MMQ com a transformação log) sobre os valores preditos de y do mesmo modelo, ambos log transformados (Buntin,2004):

$$Ln\left[\left(\hat{y}_i - \hat{y}_i\right)^2\right] = \lambda_0 + \lambda_1 Ln\left[\hat{y}_i\right] + \varphi_i$$

λ1 Indicar-nos-á que função de variância é mais apropriada aos dados, conduzindo-nos assim à família de GLM aconselhável:

$\lambda_1 \cong 0 \rightarrow$ Família gaussiana (variância constante);

$\lambda_1 \cong 1 \rightarrow$ Família Poisson (variância ∝ média);

$\lambda_1 \cong 2 \rightarrow$ Família Gamma ou Binomial (variância ∝ média2);

$\lambda_1 \cong 3 \rightarrow$ Família Inversa Gaussiana (variância ∝ média3);

Temos ainda que para o modelo em estudo, a estimativa de máxima verosimilhança de $\beta$ existe no interior do espaço paramétrico, é finita e única. [12]

### 2.2.5 Combinação de Modelos de Regressão

[12] Amaral Turkman, M.A. e Silva, G. (2000) direcciona para Wedeburn (1976).

Por vezes para ganharmos qualidade na regressão temos de fazer uma partição do domínio da variável dependente e depois fazer a regressão em cada um dos subconjuntos definidos.

Para além desta situação, poderemos ainda ter decompor a variável em estudo em duas variáveis que se combinam, resultando na variável original que se pretende explicar.

O modelo combinado, em muita literatura designado por modelo de duas partes, é necessário sempre que se trata de uma variável mista e/ou quando é definida por ramos.

***Resultado 3***- **Seja** Y uma variável aleatória definida pela seguinte expressão:

$$Y = \begin{cases} Z & \Leftarrow \quad U = 1 \\ 0 & \Leftarrow \quad U = 0 \end{cases}$$

, em qu Z e U são variáveis aleatórias e U tem distribuição de *Bernoulli*.

**Então**

- $E[Y] = E(Y|U=1)\times P[U=1] + E(Y|U=0)\times P[U=0] = E[Z]\times P[U=1] = E[Z]\times E[U]$

  $=0$, pela definição de Y

  porque U$\cap$ *Bernoulli*

- $E[Y^2] = E(Y^2|U=1)\times P[U=1] + E(Y^2|U=0))\times P[U=0] = E[Z^2]\times E[U] =$

  $= (E[Z^2] - E^2[Z] + E^2[Z])\times E[U] = (Var[Z] + E^2[Z])\times E[U]$

  $=0$, pela definição de Y

- $\boldsymbol{Var[Y]} = E[Y^2] - E^2[Y] = (Var[Z] + E^2[Z])\times E[U] - E^2[Z]\times E^2[U] =$
  $= Var[Z]\times E[U] + E^2[Z]\times E[U] - E^2[Z]\times E^2[U] =$
  $= Var[Z]\times E[U]\; E^2[Z]\times \underbrace{E[U]\times(1-E[U])} =$
  $= \boldsymbol{Var[Z]\times E[U]} + E^2[Z]\times \boldsymbol{Var[U]}$

  porque U$\cap$ Bernoulli

***Resultado 4***- **Sejam** Yi 'n' variáveis aleatórias i.i.d. como acima descrito

**Então** prova-se que

$$\frac{\bar{Y} - \mu_Y}{\sqrt{(\hat{\sigma}_Z^2 \times \mu_U + \mu_Z^2 \times \hat{\sigma}_U^2)/n}} \xrightarrow[n\to+\infty]{} N(0,1)$$

Assim temos de ter capacidade de avaliar a qualidade do modelo resultante da combinação das várias regressões, a que chamaremos de modelo combinado.

- **Qualidade do Modelo**

Perante o modelo combinado, temos de definir medidas que nos permitam tirar conclusões sobre a qualidade ou performance do modelo.

Esta avaliação pode ser feita com base em representação gráfica de estimativas vs observações ou através da análise de resíduos.

### Análise de Resíduos

Para esta análise, independentemente do tipo de regressão efectuada – linear múltipla com ou sem a variável endógena transformada ou linear generalizada – devem ser cálculadas as estimativas ( $\hat{y}_i$) na escala original e os resíduos devem ser calculados a partir dessas estimativas e das observações($y_i$) originais. Das medidas de resíduos mais frequentemente utilizadas destacam-se:

- Erro quadrático médio (RMSE de root-mean square error):

$$\boldsymbol{RMSE} = \sqrt[2]{\sum_{i=1}^{n} (\boldsymbol{y_i} - \boldsymbol{\hat{y}_i})^2 \times \boldsymbol{n^{-1}}}$$

- Erro médio absoluto(MAPE de mean absolute prediction error):

$$\boldsymbol{MAPE} = \boldsymbol{n^{-1}} \times \sum_{i=1}^{n} |\boldsymbol{y_i} - \boldsymbol{\hat{y}_i}|$$

- Coeficiente de determinação:

$$\boldsymbol{R^2} = \boldsymbol{1} - \frac{\boldsymbol{RMSE^2}}{\boldsymbol{Var}[\boldsymbol{y}]}$$

O melhor modelo é o que apresenta maior $R^2$ e menores MAPE e RMSE.

### Previsão - Intervalos de Confiança

Assumindo Y como variável aleatória conforme definida para o Resultado 3 e partindo do Resultado 4, apresentados na página anterior, podemos definir um intervalo a $(1 - \alpha/2)\%$ de confiança da seguinte forma:

$$\left( \overline{\boldsymbol{Y}} \mp \boldsymbol{\Phi^{-1}}(\boldsymbol{1} - \boldsymbol{\alpha/2}) \times \sqrt{(\boldsymbol{\hat{\sigma}_Z^2} \times \boldsymbol{\mu_U} + \boldsymbol{\mu_Z^2} \times \boldsymbol{\hat{\sigma}_U^2})/\boldsymbol{n}} \right)$$

# 3. Modelização dos custos de Ambulatório na carteira Multicare

Antes de entrar na modelização para identificação das variáveis explicativas, apresenta-se um conjunto de análises preliminares com o objectivo de selecionar, de entre das informações existentes no sistema informático da Seguradora, quais e como estas deverão ser apresentadas na determinação do custo do ambulatório,

## 3.1 Análises Preliminares

### 3.1.1 Variável Dependente

A variável dependente que se pretende estudar – CUSTO AGREGADO DAS DESPESAS DE SAÚDE EM AMBULATÓRIO – e que se apresenta na *datawarehouse* como "Valor Apresentado" é uma variável mista, com domínio em $\Re_0^+$.

| Regra de Sturges | | |
|---|---|---|
| k= | 19 | |
| $2^k$= | 524.288 | |
| dim. Amostral= | 382947 | |

| | | | | | |
|---|---|---|---|---|---|
| #Zeros = | 140.259 | | | 95,0% | 1.096,98 € |
| #Não nulos = | 242.688 | | | 96,0% | 1.186,47 € |
| Mínimo = | 2,20 € | | | 97,0% | 1.300,73 € |
| 1ºQuartil = | 94,00 € | | | 98,0% | 1.458,01 € |
| Mediana = | 225,44 € | | | 99,0% | 1.733,42 € |
| 3ºQuartil = | 466,31 € | | Outros Perentis | 99,5% | 2.056,28 € |
| Máximo | 2.295,88 € | | | 99,9% | 3.401,00 € |

O histograma resultante sugere-nos uma distribuição assimétrica:

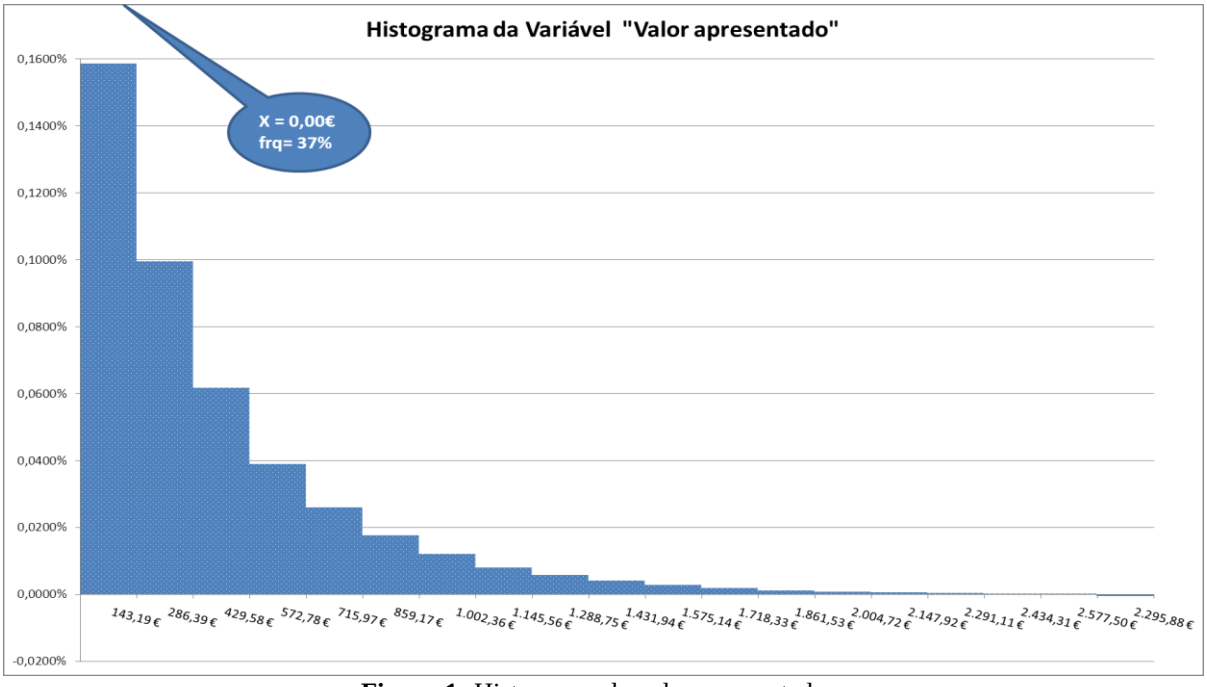



**Figura 1:** Histograma do valor apresentado

**Figura 1:** Histograma dos valores Apresentados Agregados por Pessoa Segura na cobertura de Ambulatório.

Esta variável apresenta um átomo de probabilidade no ponto zero com massa de probabilidade de 37% e que corresponde ao conjunto de Pessoas Seguras que não apresentaram despesas nesta cobertura. Nesta massa de probabilidade estão incluídas a centena (110) de Pessoas Seguras que de facto não tiveram sinistros.

### 3.1.2 Variáveis Explicativas

Existem duas variáveis explicativas dos custos da Saúde que são universalmente aceites: a Idade e o Género. Apesar desse reconhecimento, existe uma Norma Comunitária que impede de tarifar de forma diferenciada em função do género (vd,pág.10) permitindo, no entanto, que todas as restantes medidas de risco da atividade seguradora, tais como Provisões Técnicas, possam utilizar esse parâmetro de diferenciação.

A amostra que vamos trabalhar tem, no que respeita a essas características, a seguinte composição:

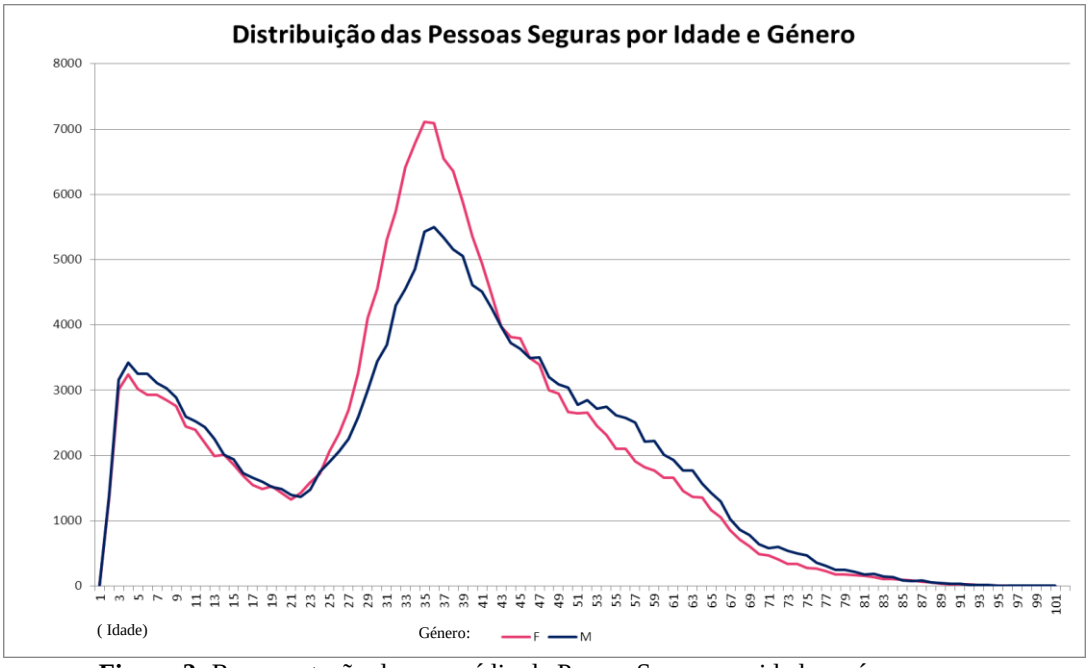



**Figura 2:** Representação do vap médio da Pessoa Segura por idade e género

Num total de trezentas e oitenta e três mil (382.947) pessoas seguras, cerca de cinquenta e um por cento (195.907 pessoas seguras) são do género feminino.

Para a amostra - cobertura - foram selecionadas como potenciais variáveis explicativas as seguintes características: Tipo de Seguro (Individual ou Grupo), Grupo de Produto, Subgrupo de Produto, Família de Produto, Data do Plano, Limite da Despesa (Capital Seguro), %Comparticipação, Idade da Pessoa Segura, Parentesco, Localidade Postal, Concelho, Distrito e Zona Multicare.

- **Tipo de Seguro**

## Seguros de Grupo vs Seguros Individuais

Foi feita uma representação do perfil de custo médio e do perfil do número médio de sinistros por idade, com as idades onde se verificam mais de 100 indivíduos em qualquer dos tipos de seguro, para verificar se existe razão para pensar que esta varável pode ser explicativa da variabilidade da curva de indemnizações:

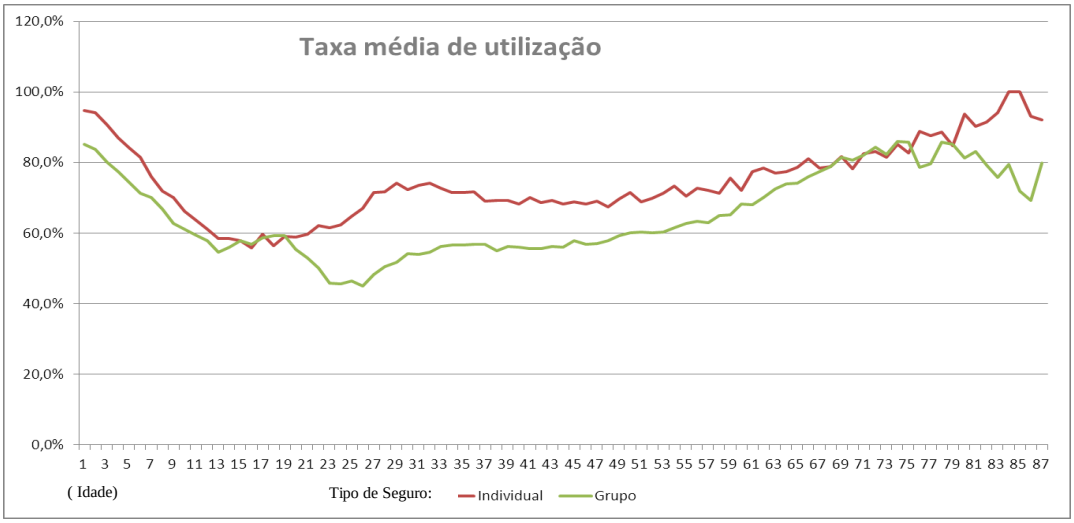



**Figura 3:** Representação da Taxa média de utilização por idade e tipo de seguro

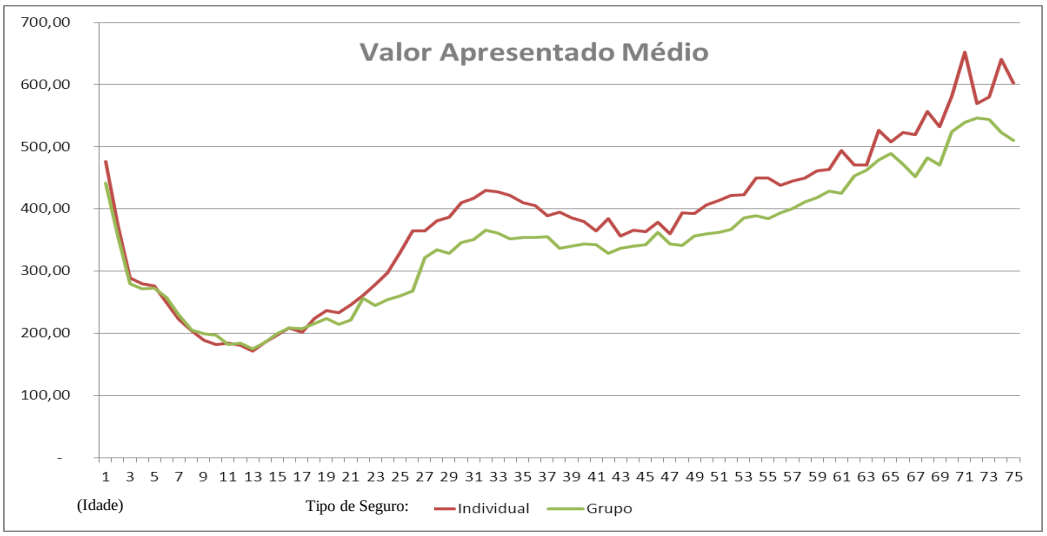



**Figura 4:** Representação do vap médio da Pessoa Segura por tipo de seguro

O Tipo de Seguro é uma variável que esperamos tenha algum poder explicativo, pois identifica contratos com processos de tarifação e de seleção de risco muito distintos, veja-se a anti-seleção da carteira individual sugerida nas figuras acima.

- **Tipo de Produto**

### Grupo de Produto, Subgrupo de Produto, Família de Produto

Estas variáveis são, todas elas, variáveis qualitativas, o que significa que ao entrarem no modelo de regressão serão necessariamente convertidas em variáveis dummy. Assim teremos de verificar se se mantém o "interesse" direto nestas variáveis ou se deveremos utilizar qualquer transformação delas.

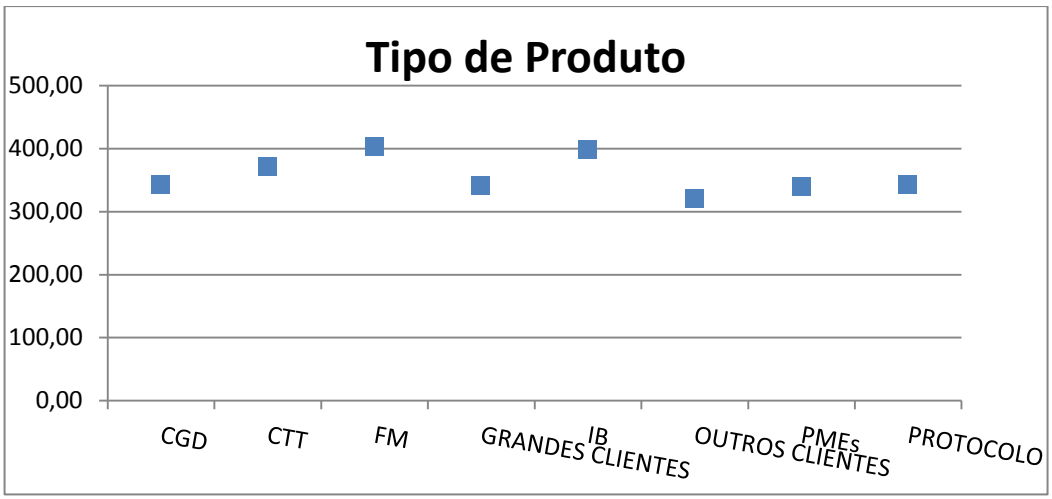



**Figura 5:** Representação do vap médio da Pessoa Segura por tipo de produto

Uma vez se tratarem de custos médios com alguma proximidade, optámos por fazer uma análise de *clusters* sobre esta característica com base nas duas variáveis "vap" e "ocorr" para se verificar a possibilidade, ainda assim, em agrupar por forma a reduzir o número de classes, uma vez que algumas delas apresentam características semelhantes:

| | Tipo de Produto | Pss.Seguras | Cl.Utilizadores | Tx.Utilização | Vl Apresentado |
|---|---|---|---|---|---|
| 1 | CGD | 47.515 | 32.416 | 68,2% | 342,91 € |
| 2 | CTT | 653 | 548 | 83,9% | 371,55 € |
| 3 | FM | 24.346 | 19.068 | 78,3% | 402,94 € |
| 4 | GRANDES CLIENTES | 196.438 | 116.571 | 59,3% | 340,71 € |
| 5 | IB | 27.382 | 20.577 | 75,1% | 398,88 € |
| 6 | OUTROS CLIENTES | 53.888 | 31.407 | 58,3% | 320,45 € |
| 7 | PMEs | 22.157 | 14.855 | 67,0% | 340,30 € |
| 8 | PROTOCOLO | 10.559 | 7.245 | 68,6% | 342,82 € |

**Tabela 3:** Vap médio e tx. média de utilização por tipo de produto

Nas restantes características que estão incluídas no sistema técnico de gestão da Seguradora sobre a caracterização do produto e que se representa de seguida, para aqueles que surgem com uma tão grande dispersão, optou-se pela respectiva substituição.

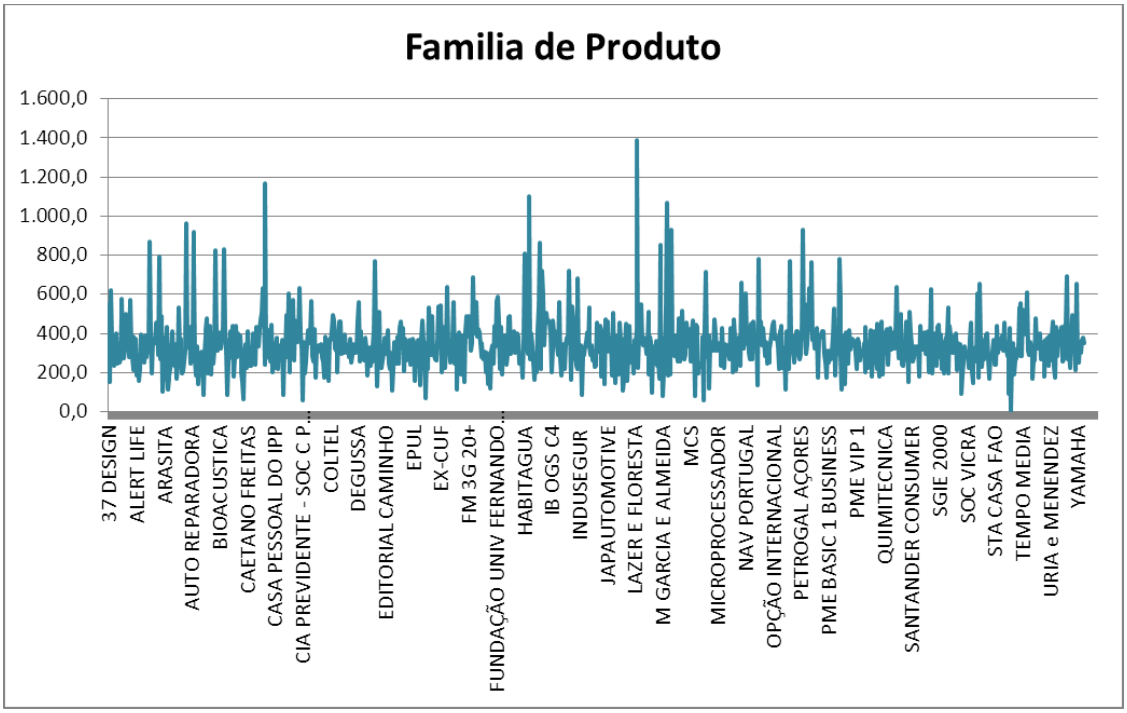



**Figura 6:** Representação do vap médio da pessoa segura por família de produtos

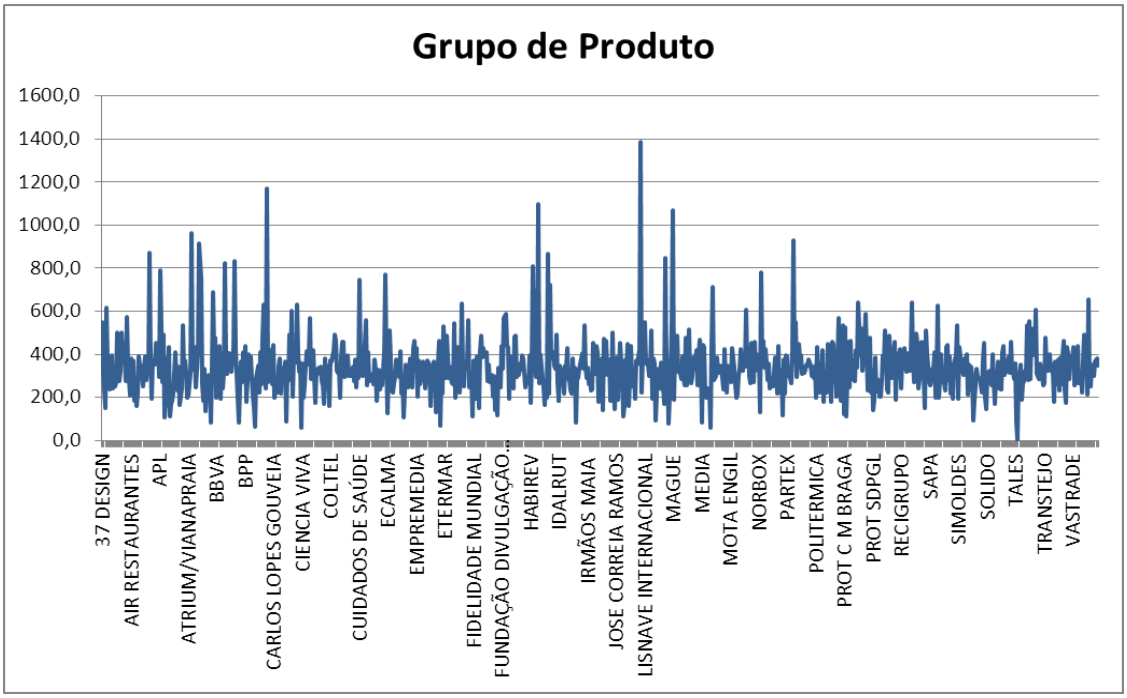



**Figura 7:** Representação do vap médio da pessoa segura por grupo de produto

Uma alternativa possível de abordagem a estas últimas duas variáveis, será o número de coberturas, talvez a característica mais determinante do produto. O número de coberturas num Seguro de Saúde da Multicare, e da maioria das operadoras no mercado Português, traduz quais são essas as coberturas já que estas têm, na prática, regras implícitas de precedência.

Assim, definamos

"Nº Coberturas do Produto" =
- 1 <= Apenas inclui Internamento
- 2 <= Inclui Internamento e Ambulatório
- 3 <= Inclui Internamento, Ambulatório e Estomatologia
- 4 <= Inclui mais coberturas

Fazendo agora uma análise de dispersão dos valores apresentados da amostra em função destas novas classes, verifica-se que esta pode ser uma variável explicativa, pois apesar de se construir a tarifa por cobertura, os valores apresentados da mesma diferem quando inseridos em produtos mais ou menos completos.

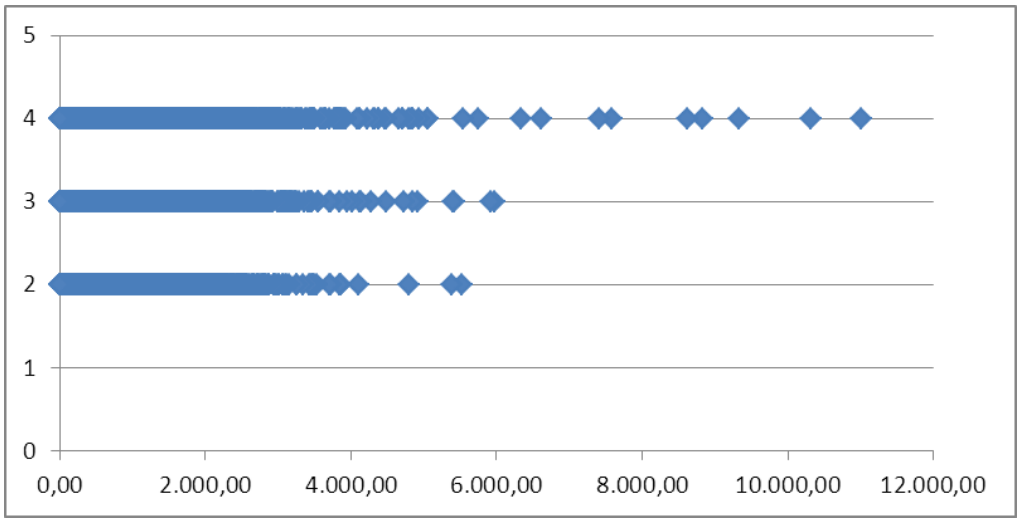



**Figura 8:** Representação do vap médio da pessoa segura por grupo de produto

- **Estrutura do Produto**

### Limite da despesa (Capital Seguro), Percentagem de Comparticipação

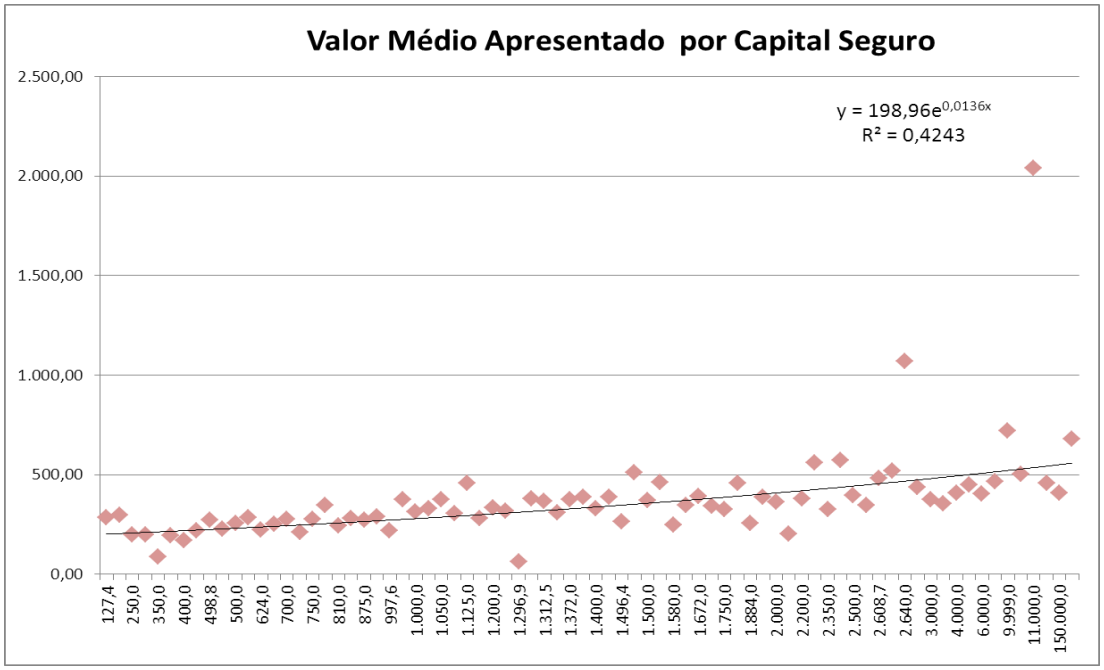



**Figura 9:** Representação do vap médio ~ Capital Seguro da cobertura

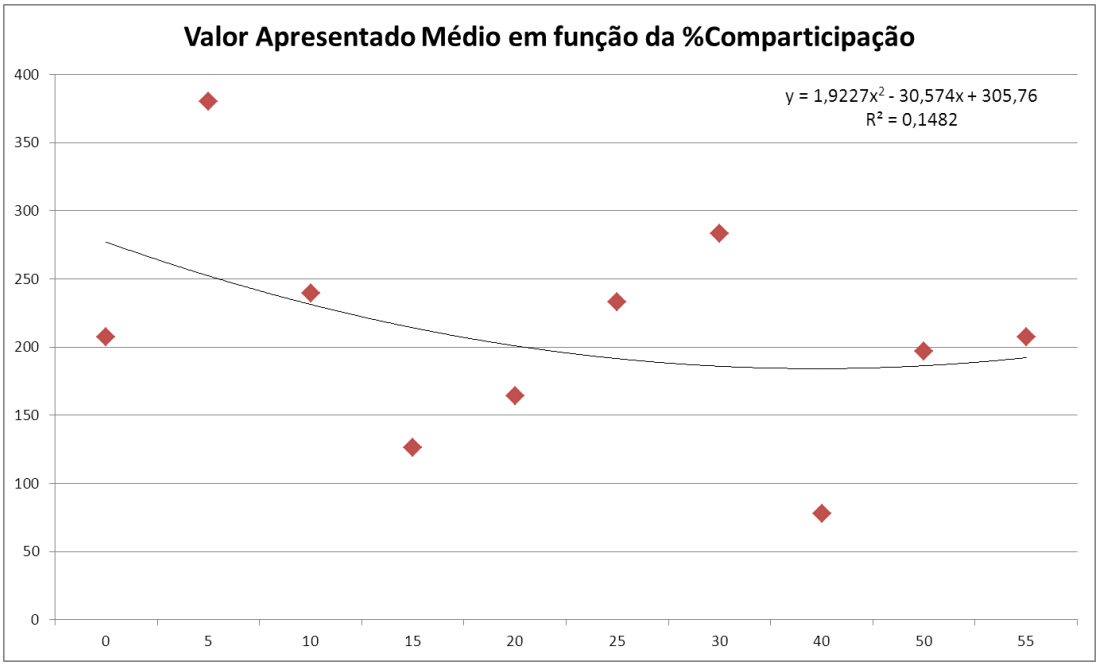



**Figura 10:** Representação do vap médio da Pessoa Segura por % comparticipação do cliente

Aparentemente parece não haver uma correlação muito forte entre qualquer das variáveis - capital seguro e percentagem de comparticipação - e a variável dependente, mas como qualquer destas variáveis é quantitativa vamos incluí-las na regressão, sem qualquer tratamento prévio.

- **Caracterização da Pessoa Segura**

### Idade, Sexo e Parentesco

A Idade, como variável explicativa, confirma-se como sendo de primordial importância:

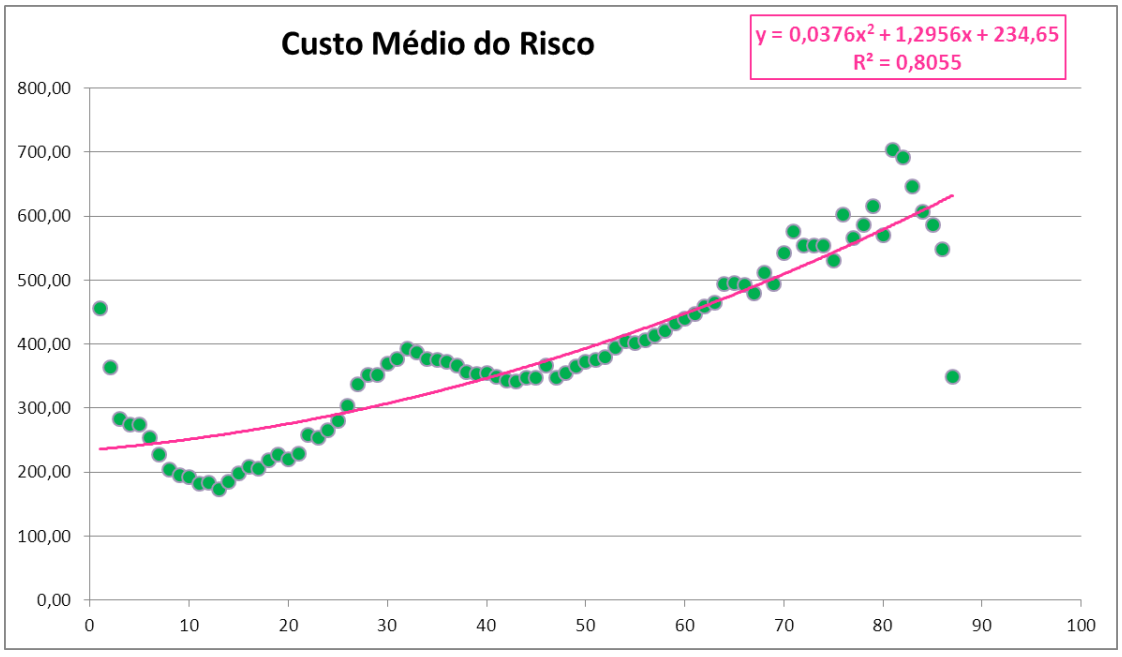



**Figura 11:** Representação do vap médio da Pessoa segura em função da idade

Aparentemente, nesta medida, existem três partições que se caracterizam pela fase de crescimento – dos 0 aos 18 anos – fase de procriação – dos 19 aos 40 anos – e de maturidade – a partir dos 41 anos – com diferentes impactos no valor médio apresentado. Assim levanta-se a questão: fará sentido construir três modelos de custo independentes para cada fase?

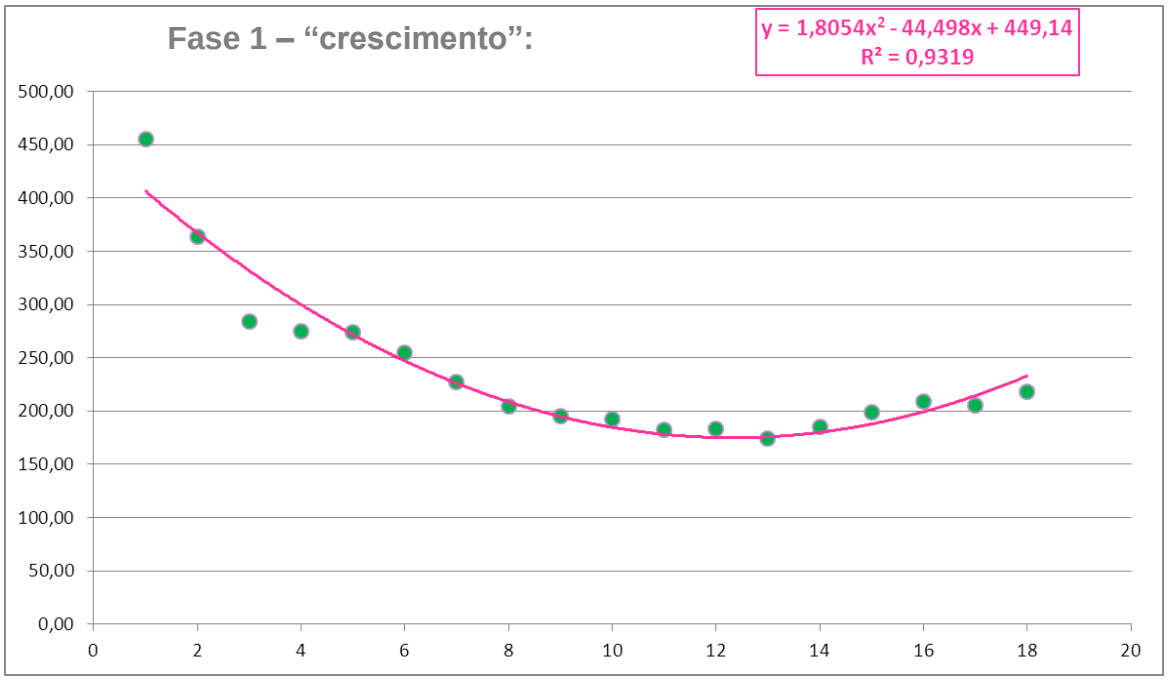



**Figura 12:** Representação do vap médio da Pessoa Segura em função da idade

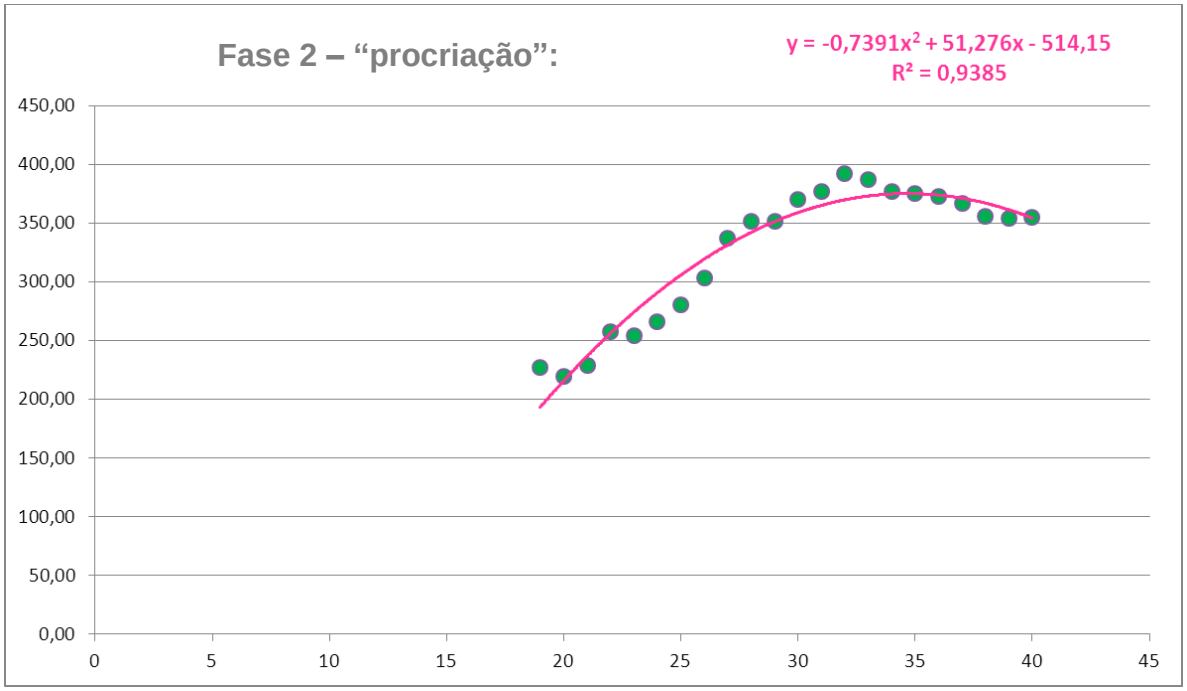



**Figura 13:** vap médio da Pessoa Segura em função da idade na fase "procriação"

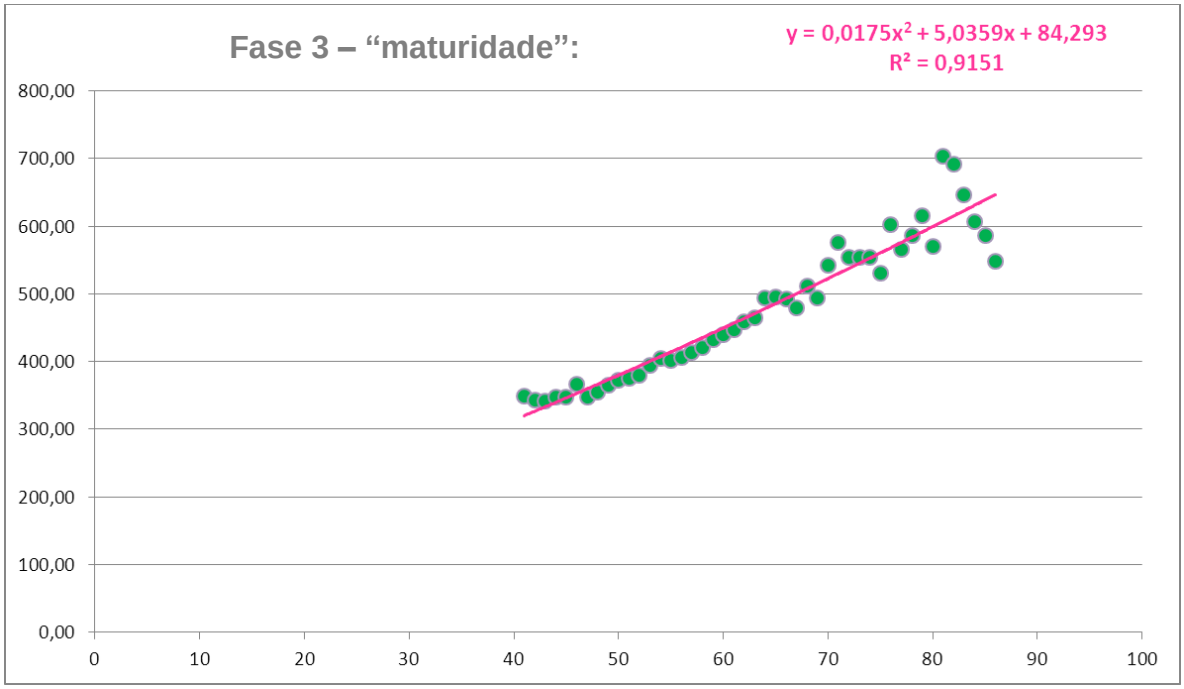



**Figura 14:** Vap médio da Pessoa Segura em função da idade na fase "maturidade"

Nos estudos desenvolvidos foram ensaiados modelos para todas as idades – adiante designados como "modelo global" – e para as diversas faixas etárias, ainda que com alguns ajustamentos face a estas que aqui se apresentam por revelarem melhores resultados.

Como já foi referido no início desta secção – pág.35 – e na secção "Motivação e Objetivos" – pág.10 – o Género é uma variável muito reconhecida e caracterizadora da população segura, resta-nos então confirmar se em termos de custo do risco ela é, como se espera, significativa:

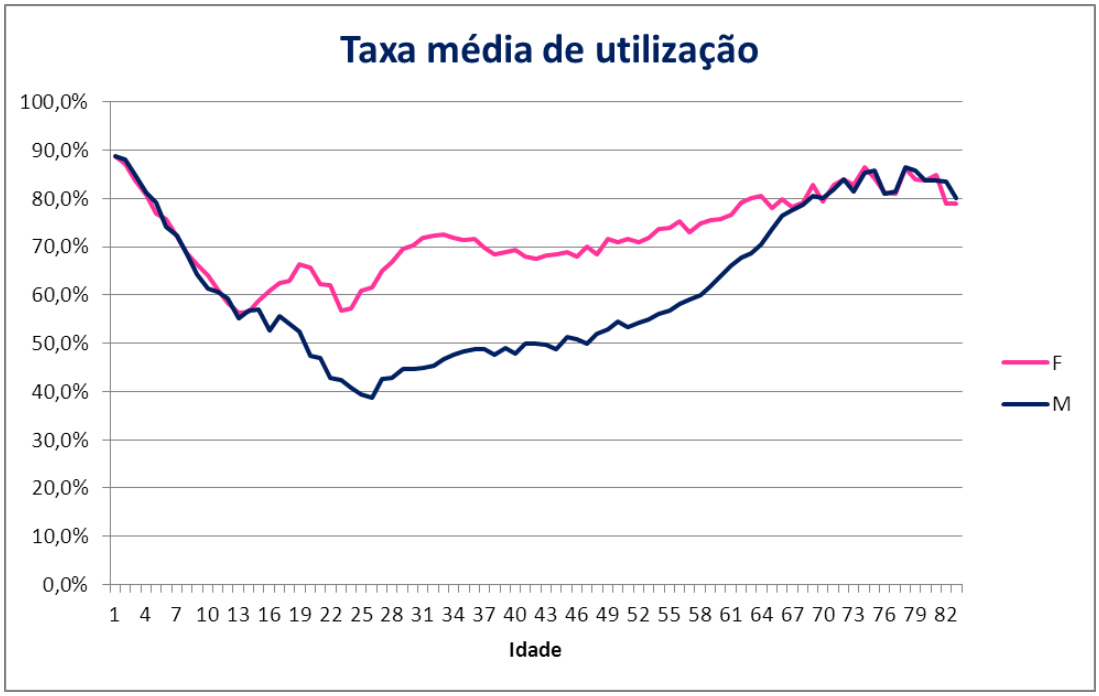



**Figura 15:** Taxa média de utilização por idade e género

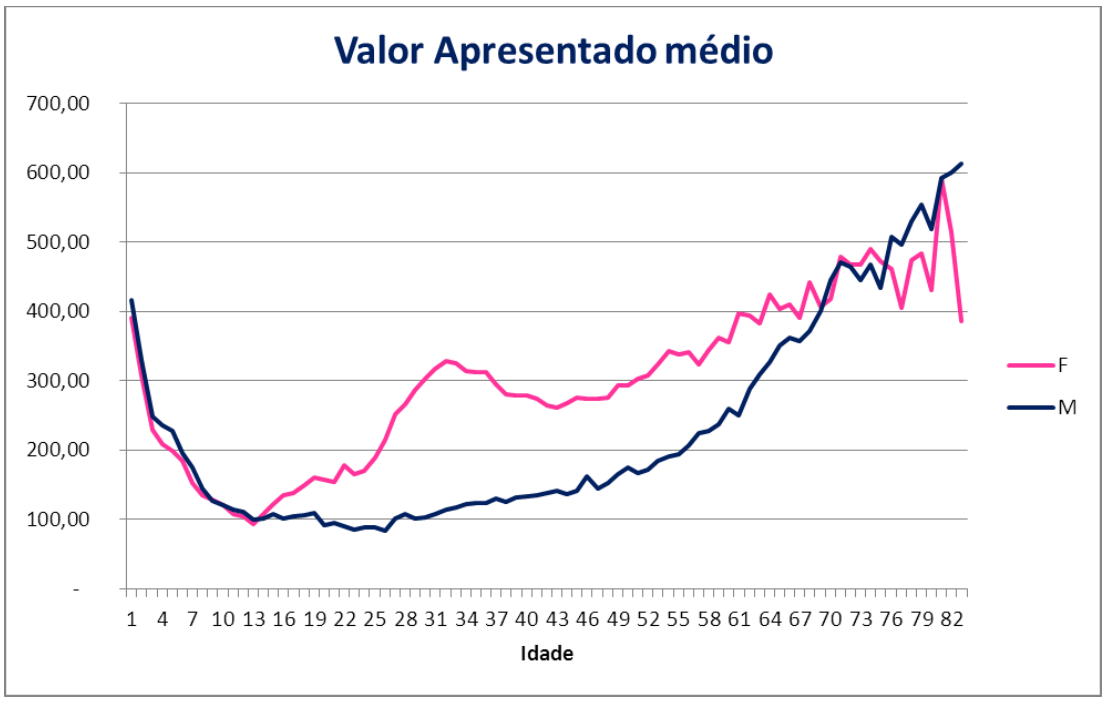



**Figura 16:** Vap médio por idade e género

Como se pode observar, quer em termos de utilização, quer em termos de custo médio do risco, estamos perante uma variável explicativa significativa.

O Parentesco é outra das variáveis exógenas que é tradicionalmente reconhecida como uma variável explicativa do custo do risco:

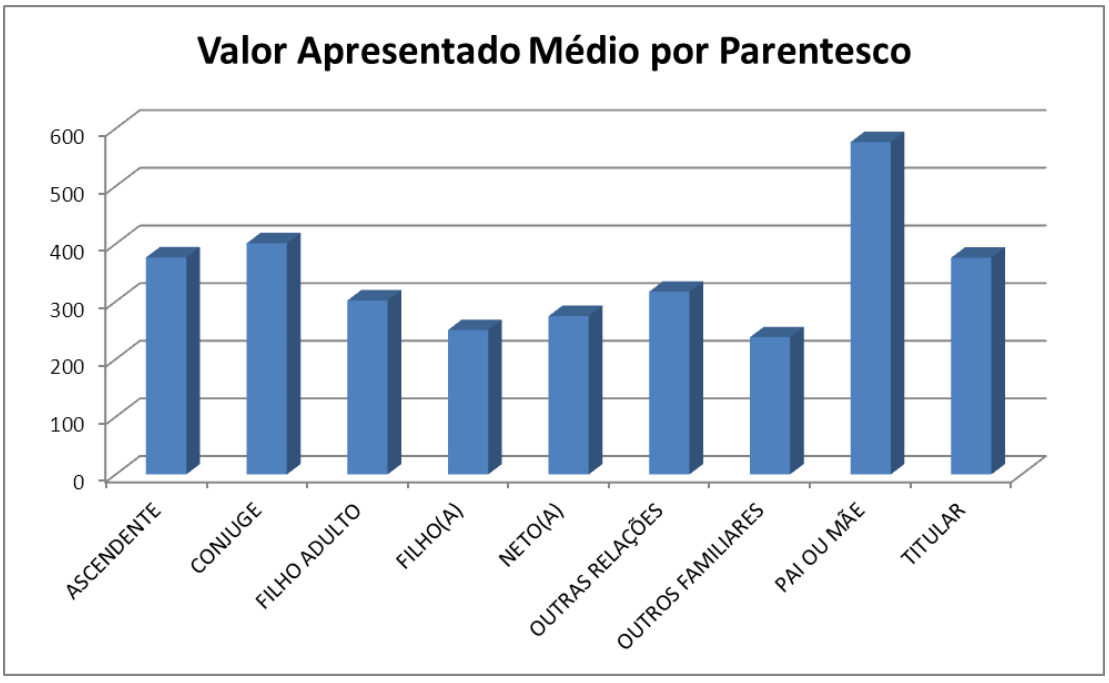



**Figura 17:** Representação do vap médio da Pessoa Segura por parentesco

De facto parece haver diferenças significativas entre alguns dos parentescos, não sendo todos significativamente diferentes.

Quisemos verificar se se tratava, apenas, de uma influência indireta, isto é, se o parentesco poderia ajudar a explicar o custo com sinistros por este estar correlacionado com a idade da Pessoa Segura:

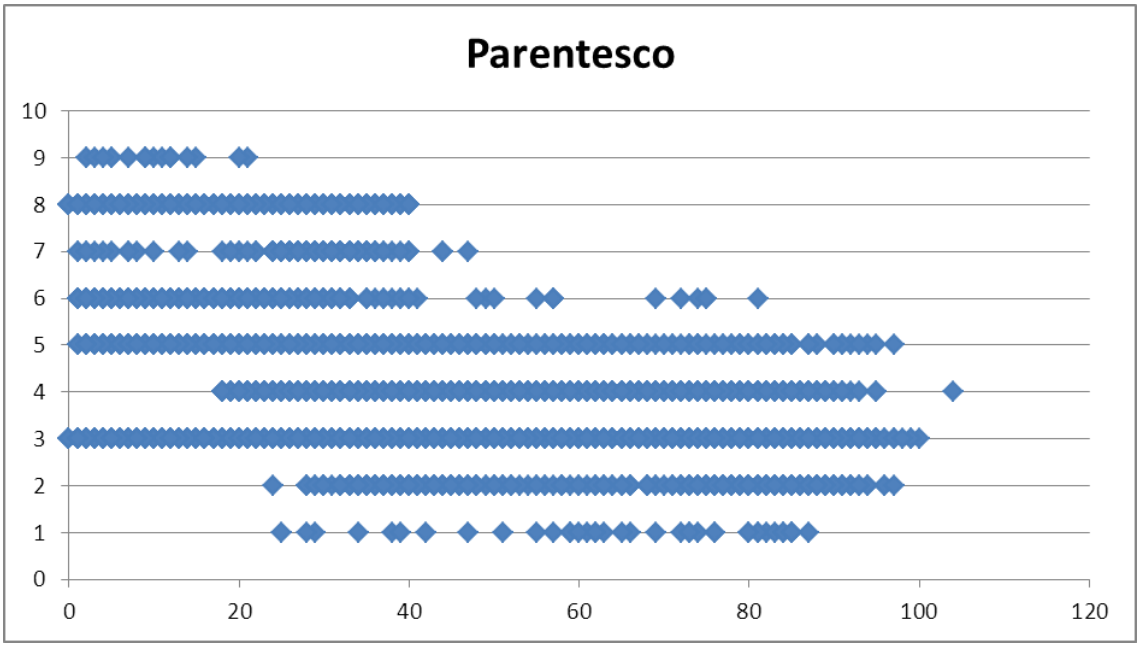



**Figura 18:** Representação das idades das pessoa Seguras e do parentesco com o titular da apólice
**Legenda –** 1:Ascendente; 2:Pai ou Mãe; 3:Titular; 4:Conjuge; 5:Outras Relações; 6:Outros Familiares;7:Filho Adulto; 8:Filho(a); 9:Neto(a)

O parentesco, quando se trata de descendência ou de ascendência direta, pressupõe um diferencial de idades significativo, como se pode observar nos "ascendentes", "pai ou mãe", "filho", "filho adulto" ou "neto". O cônjuge também só surge em idades adultas. Mas para os restantes parentes, incluindo titulares, existem de todas as idades. Assim valerá a pena introduzir esta característica como variável explicativa.

Ora ela é também uma variável qualitativa, tornando-se, por isso, alvo de uma análise de *Clusters* que referiremos em secção própria.

Quisemos perceber algumas das associações e dissociações baseadas na relação idade~parentesco. Fomos por isso observar as duas dimensões destas classes - a idade média e o custo médio apresentado:

| | Parentesco | Pss.Seguras | Idade média | Cl.Utilizadores | Tx.Utilização | Vl Apresentado |
|---|---|---|---|---|---|---|
| 1 | ASCENDENTE | 39 | 64,4 | 30 | 76,92% | 376,27 € |
| 2 | PAI OU MÃE | 455 | 64,3 | 372 | 81,76% | 471,07 € |
| 3 | TITULAR | 233.183 | 40,8 | 140.783 | 60,37% | 226,87 € |
| 4 | CONJUGE | 58.451 | 45,5 | 39.581 | 67,72% | 271,38 € |
| 5 | OUTRAS RELAÇÕES | 664 | 30,5 | 472 | 71,08% | 225,60 € |
| 6 | OUTROS FAMILIARES | 1.263 | 10,5 | 816 | 64,61% | 153,68 € |
| 7 | FILHO ADULTO | 485 | 27,7 | 311 | 64,12% | 193,00 € |
| 8 | FILHO(A) | 88.392 | 11,1 | 60.310 | 68,23% | 170,80 € |
| 9 | NETO(A) | 19 | 9,6 | 14 | 73,68% | 202,32 € |

**Tabela 4:** Alguns indicadores de gestão das classes de Pessoas Seguras por parentesco

Observando os vap médios, verifica-se que a idade não explica todas as variações nos custos médio: de notar que os netos sendo os que têm a idade média mais baixa não são os que apresentam menor custo e o filho adulto, com idade média de 28 anos, apresenta uma taxa de utilização inferior à do filho (criança), que apresenta uma idade média de 11 anos, pelo que se conclui que o parentesco não deve ser abandonado como potencial variável explicativa.

- **Variável Geográfica**

### Localidade Postal, Concelho, Distrito e Zona Multicare

Todas estas variáveis são qualitativas, por isso todas terão de ser substituídas por variáveis dummy. A Localidade Postal pode ser agrupada em Concelhos, Distritos ou Zona Multicare. Devido ao elevado número de localidades e concelhos existentes no país, o que, como já referimos anteriormente, tornaria o modelo ingerível, optámos por analisar unicamente as classificações que conduzem a um menor número de classes: Distrito e Zona Multicare.

Apesar desse número de classes ser menor, optou-se ainda por fazer uma análise de *clusters* que permitisse reduzir o número de classes em qualquer uma das divisões das áreas geográficas:

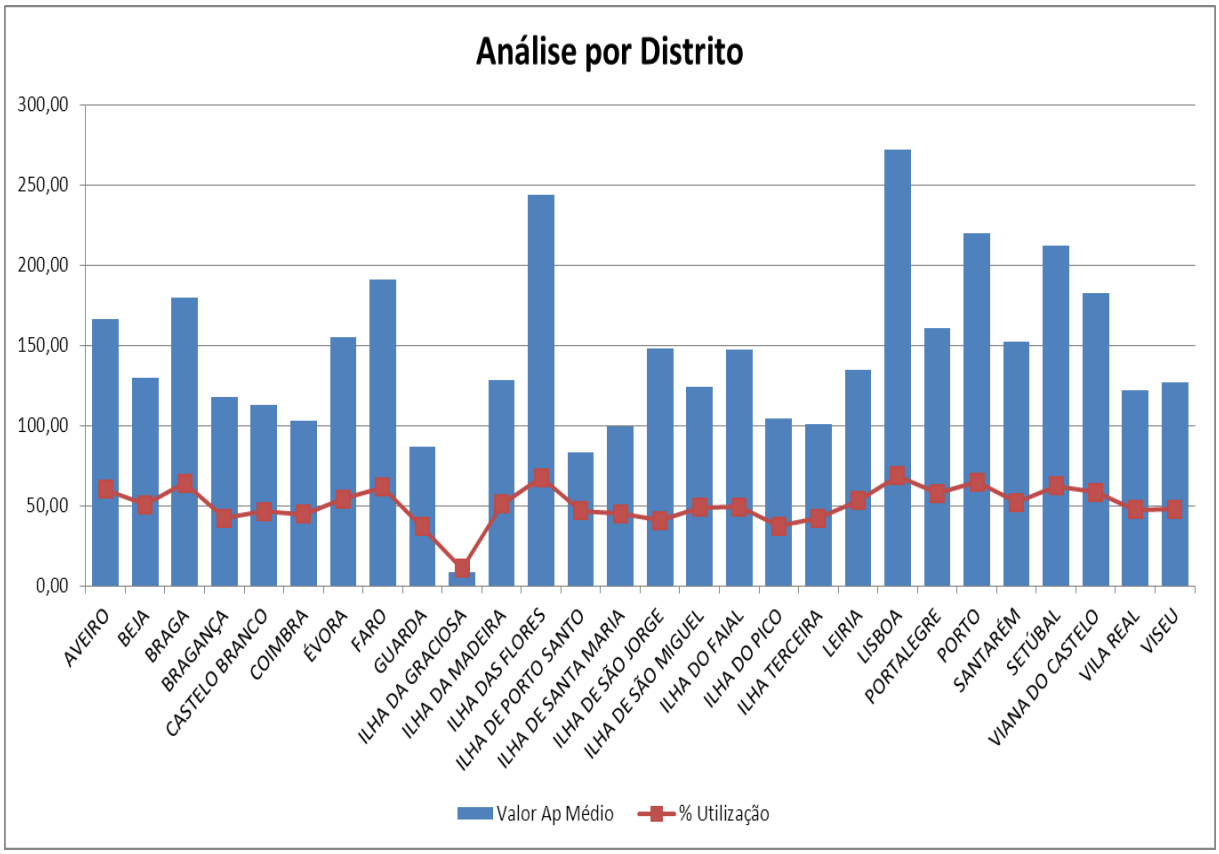



**Figura 19:** Representação do vap médio do coletivo Pessoas Seguras de cada Distrito

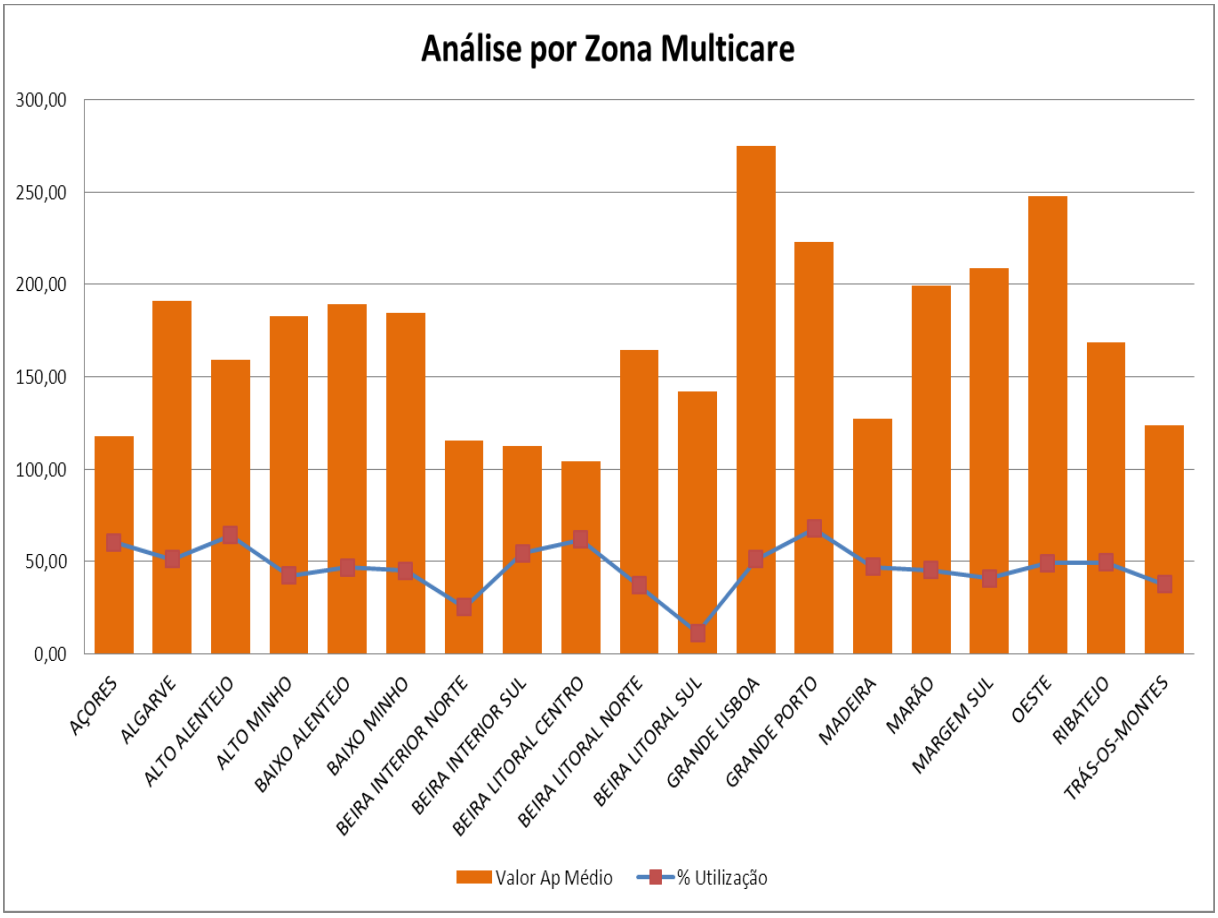



**Figura 20:** Representação do vap médio do coletivo Pessoas Seguras por Zona Multicare

O Distrito e a Zona Multicare são classificações alternativas. Esta segunda classificação será por ventura mais ajustada ao risco segurador, já que, como o nome indica, se trata de uma classificação interna que resulta da actividade da Seguradora. Esta é a situação que se vai analisar quando se ensaiarem os modelos de regressão.

Caso a Zona Multicare não revele que produz claramente um melhor ajustamento do risco, então optaremos pelo Distrito, por se tratar de uma divisão oficial e indiscutível. A rede comercial tem já referido, por diversas vezes, que existe uma dificuldade prática no reconhecimento desta divisão do território e que, portanto, só deverá continuar a ser usada caso traga um claro benefício.

## 3.2 Modelização dos Custos de Ambulatório

### 3.2.1 Análise de *Clusters*

Na base de dados das apólices com cobertura de Ambulatório da Multicare foram identificadas um conjunto de quinze variáveis com eventual possibilidade de alguma forma serem explicativas do risco. Nesse conjunto de variáveis foram identificadas nove do tipo qualitativo com vários níveis de classificação, conforme se mostra na tabela abaixo:

| Variável Qualitativa | # Níveis de Classificação |
|---|---|
| Tipo de Seguro | 2 |
| Tipo de Produto | 8 |
| Família de Produto | 987 |
| Grupo de Produto | 872 |
| Parentesco | 9 |
| Distrito | 27 |
| Zona Multicare | 21 |

**Tabela 5:** Níveis de classificação das variáveis qualitativas

Ora qualquer uma destas variáveis para ser utilizada na explicação da variável resposta de um Modelo de Regressão (custo do risco) careceria de ser convertida em tantas variáveis *dummy* quantos o número de níveis menos um o que significava trabalhar com quase duas mil (1.919) variáveis, para além das restantes seis variáveis quantitativas.

Assim eliminámos a localidade Postal e Concelho, integrando os respetivos agrupamentos em Distritos e mantivemos, como alternativa, a Zona Multicare, o que como já referimos na secção 3.1.2 .

Fizemos o mesmo tipo de opção quanto à classificação dos produtos, acrescentando-lhe um campo com o número de coberturas uma vez que seria uma das informações que perderíamos ao dispensar a família de Produto.

No que respeita à variável "Tipo de Seguro", uma vez que se trata de uma classificação binária, não requer qualquer tipo de tratamento prévio.

Com as restantes variáveis, a saber "Tipo de Produto", "Parentesco", "Distrito" e "Zona Multicare" fez-se uma análise de *Clusters* com base no "vap", na "ocorr" e no "n" (número de pessoas em risco) para reduzir o número de dummies , mas tentando perder o mínimo possível da capacidade de explicação do risco.

Em todas as variáveis analisadas, usou-se o mesmo processo:

1. Determinação do nº de *clusters* pelo processo "baseado na variância explicada pela classificação"(conforme apresentado em 2.1.2), por forma a garantir que a variabilidade que se perde, dentro dos *clusters*, é inferior a 10%
2. Classificação em *clusters* pelo método hierárquico que produzir melhores resultados. A verificação é feita através do índice de cofonética
3. Classificação ótima pelo método do *kmeans*, a qual foi obtida no passo 1, para o número de *clusters* selecionado.

- **Tipo de Produto**

Para esta variável, o número de *clusters* determinado foi três, o que garante a explicação da variabilidade da amostra em 90,5%. Os tipos de Produtos foram a seguir agrupados em redes tradicionais, grandes clientes e outros.

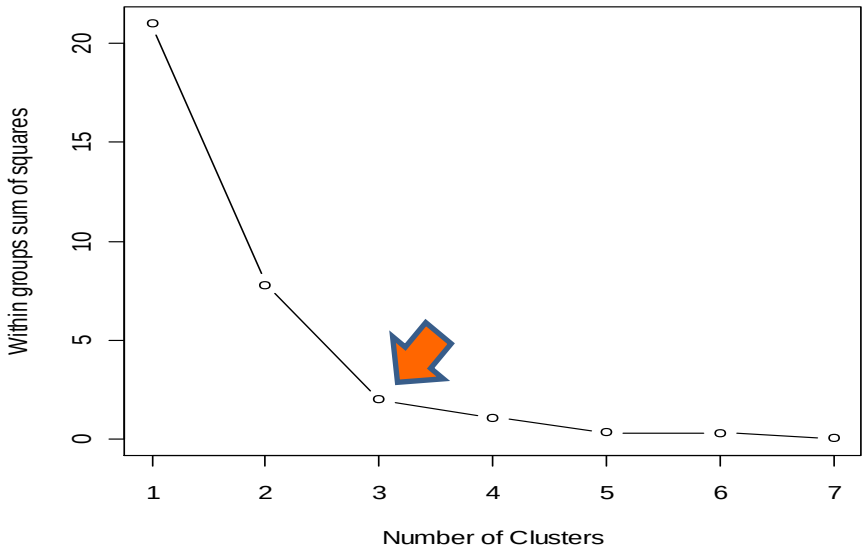



**Figura 21:** Variabilidade perdida com a classificação (pelo k-means) em função do número de clusters para o tipo de produto

O método hierárquico *average*, com a distância euclidiana, foi o que apresentou melhor índice do cofonética. A classificação resultou idêntica à definida pelo *kmeans*.

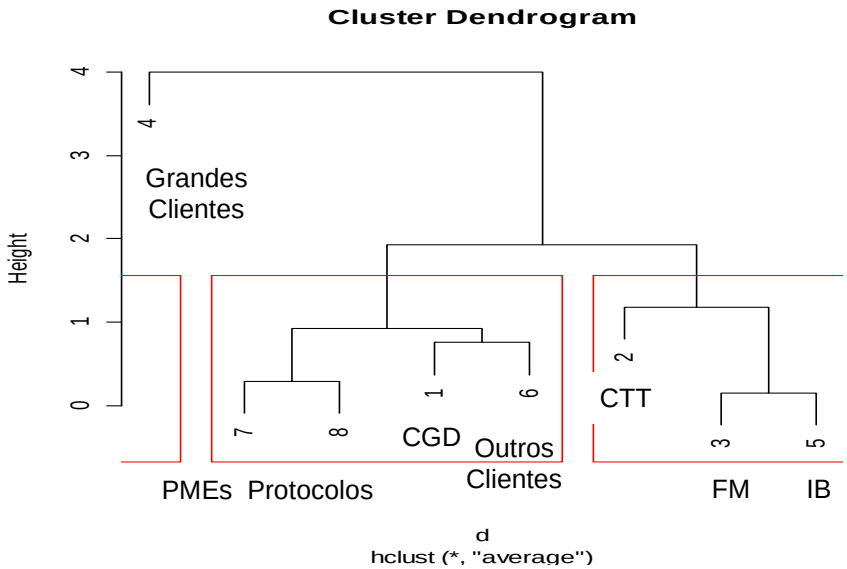



**Figura 22:** Dendrograma resultante da classificação hierárquica (average) do tipo de produto[13]

[13] Legenda completa no Anexo 4

- **Parentesco**

Para o parentesco da pessoa segura com o titular, o processo ditou quatro *clusters* conseguindo explicar 93,4% da variabilidade da amostra.

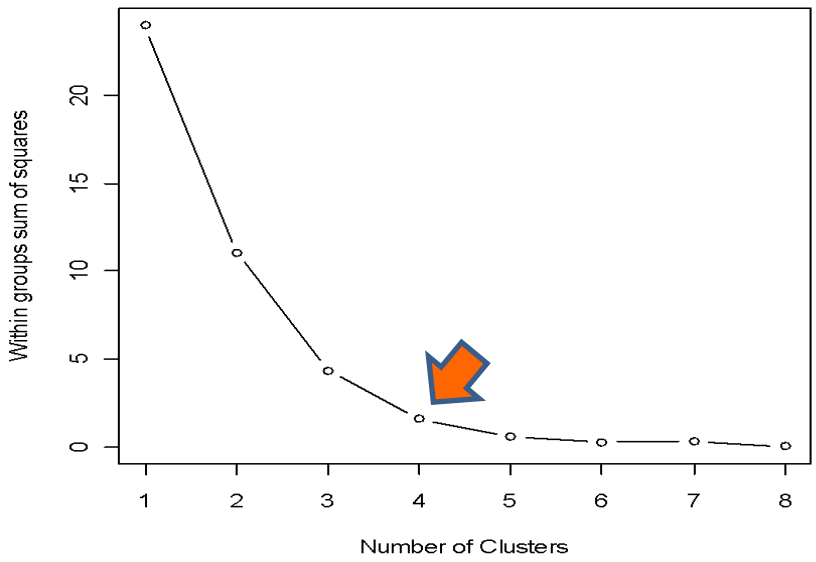



**Figura 23:** Variabilidade perdida com a classificação (pelo k-means) em função do número de clusters para o parentesco

A definição dos *clusters*, que resultou idêntica para o *kmeans* e para o método hierárquico *average*, que determinou o melhor índice de cofonética (0,933), foi a separação de todos os outros do titular, do agregado familiar, em sentido estrito, e dos mais velhos (pai, mãe ou outros ascendentes).

Esta classificação suportou-se nas três variáveis "vap", "ocorr" e "n", não estando exclusivamente ligada com a idade do parente:

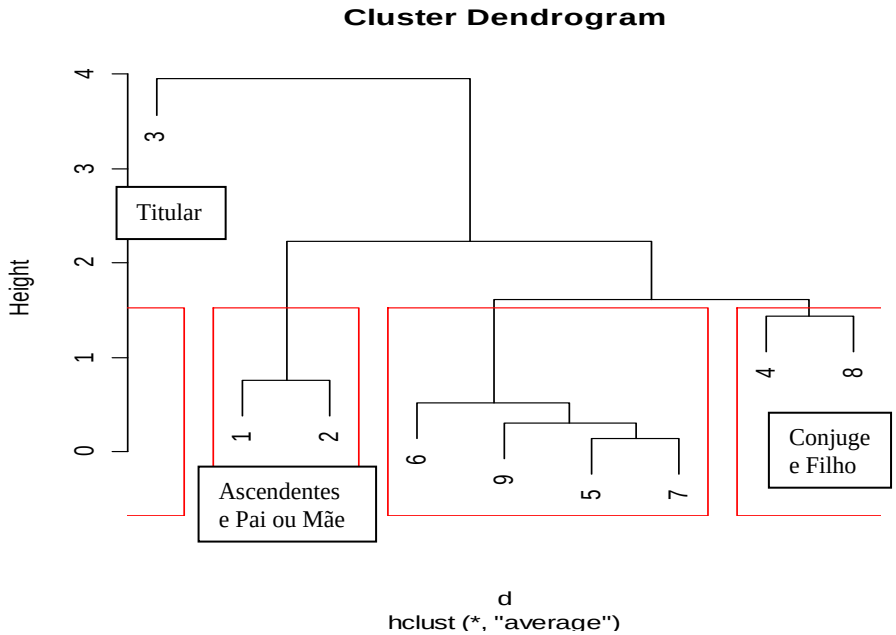



**Figura 24:** Dendrograma resultante da classificação hierárquica (average) de parentesco

- **Distrito e Zona Multicare**

Como já foi referido, estas duas variáveis classificam o mesmo vector – residência – e dessa forma são sempre alternativas. Não obstante, fizemos ambas as classificações.

Na classificação por Distrito o *kmeans* ditou uma classificação em 5 *clusters*, explicando assim 94,1% da variabilidade da amostra.

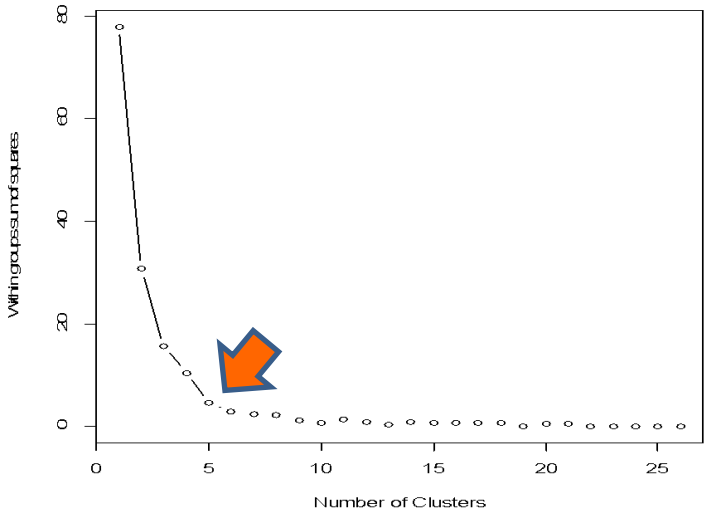



**Figura 25:** Variabilidade perdida com a classificação (pelo k-means) em função do número de clusters para o Distrito

Os *clusters* obtidos pelo *kmeans* e pelo modelo hierárquico, para esta variável, ditaram classificações diferentes:

| *Clusters pelo Kmeans* | |
|---|---|
| ***Cluster*** | ***Distrito*** |
| **1** | 8, 11, 14, 25 |
| **2** | 20 |
| **3** | 5,6,9,12,13,18 |
| **4** | 1, 2, 3, 4, 7, 10, 15, 16, 17, 19, 21, 23, 26, 27 |
| **5** | 22, 24 |

**Tabela 6:** Clusters resultantes da classificação pelo k-means dos Distritos[14]

[14] Descodificação dos Distritos e Zonas Multicare no Anexo 4

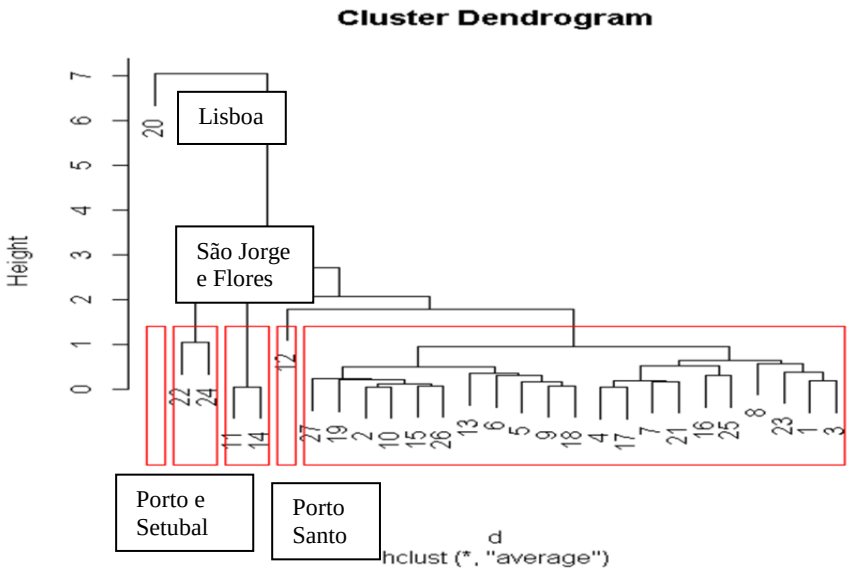



**Figura 26:** Dendrograma resultante da classificação hierárquica (average) do Distrito[14]

Finalmente, no que diz respeito à Zona Multicare, a situação foi bastante semelhante em termos do comportamento distinto dos dois métodos selecionados: O número ótimo de *clusters* foi de quatro (4), com uma variabilidade explicada de 92%:

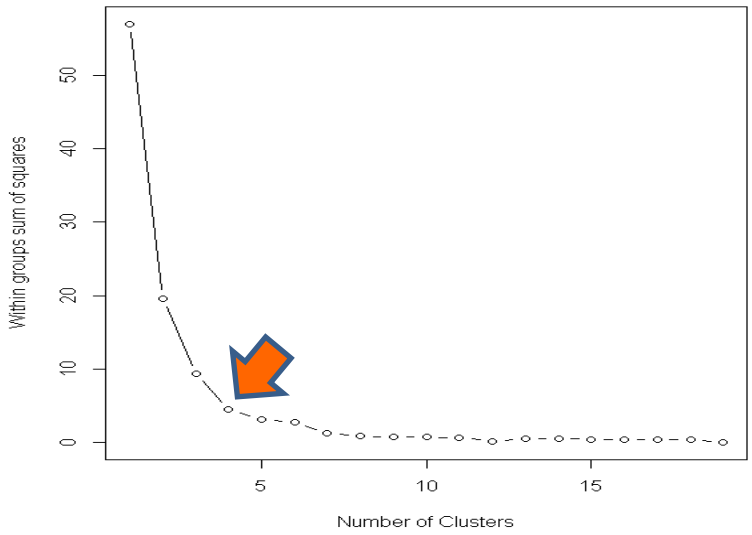



**Figura 27:** Variabilidade perdida com a classificação (pelo k-means) em função do número de clusters para a Zona Multicare

E as classificações foram:

| *Clusters pelo Kmeans* | |
|---|---|
| ***Cluster*** | ***Zona Multicare*** |
| **1** | 12 |
| **2** | 13, 16, 17 |
| **3** | 1, 7, 8, 9, 11, 14, 19 |
| **4** | 2, 3, 4, 5. 6, 10, 15, 18 |

**Tabela 7:** Clusters resultantes da classificação pelo k-means das Zonas Multicare [14]

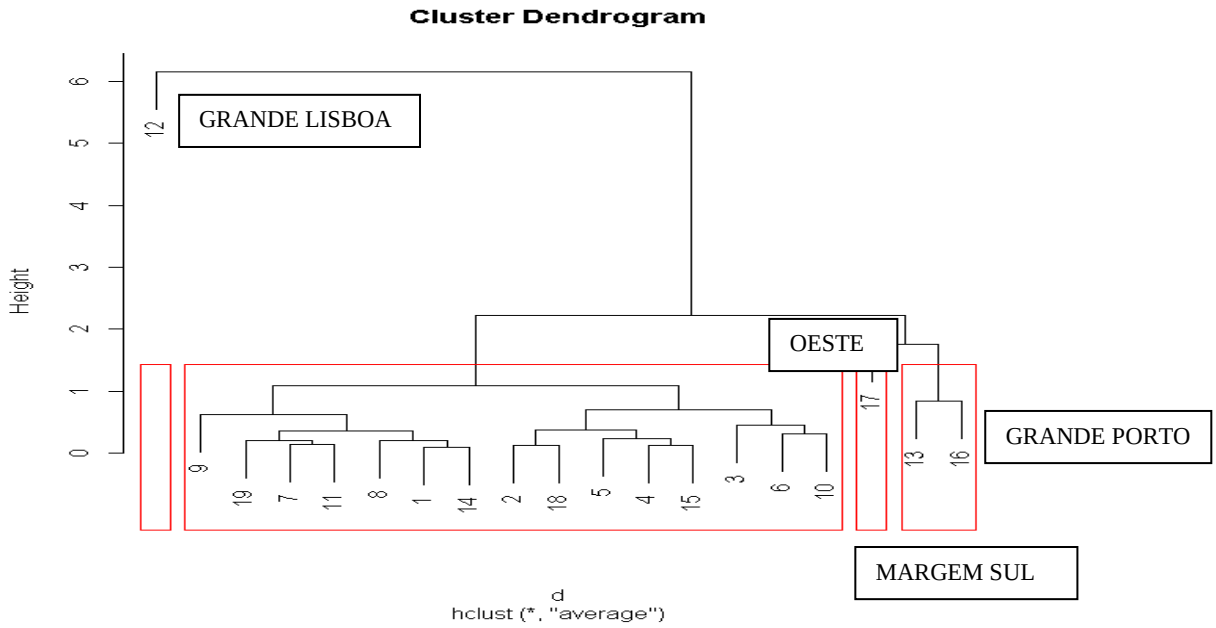



**Figura 28:** Dendrograma resultante da classificação hierárquica (average) da Zona Multicare

### Selecção dos *Clusters*

No que respeita ao tipo de produto e ao grau de parentesco qualquer das técnicas conduziu aos mesmos *clusters*:

- ✓ no parentesco (titular, pai ou mãe ou ascendente, cônjuge ou filho e os restantes parentes)
- ✓ no tipo de seguro a classificação foi: redes tradicionais ou CTT, grandes clientes e outros.

Falta então optar por usar os Distritos ou as Zonas Multicare, caso estas aumentem significativamente a variabilidade explicada do modelo, e ainda, escolher qual das classificações de cada uma destas variáveis vamos usar, já que os modelos hierárquicos e os não hierárquicos conduziram a *clusters* distintos.

A decisão será tomada com base nos resultados do primeiro modelo que se ensaiar.

## 3.2.2 Modelo de Custo

### 3.2.2.1 Seleção do Modelo

No modelo de valor Apresentado dos clientes utilizadores da cobertura, modelo que assenta numa distribuição continua, positiva e assimétrica, o histograma sugere a utilização de uma Gamma ou uma Lognormal. Habitualmente, a distribuição das indemnizações individuais tem esse comportamento, pelo que o modelo de Regressão Linear Generalizada que poderá ser associado é o modelo de uma distribuição Gamma ou o modelo de Regressão Linear com a variável endógena transformada.

Para apoio à selecção do modelo, construiu-se um histograma para comparação dos dados da amostra com as possíveis distribuições da variável dependente:

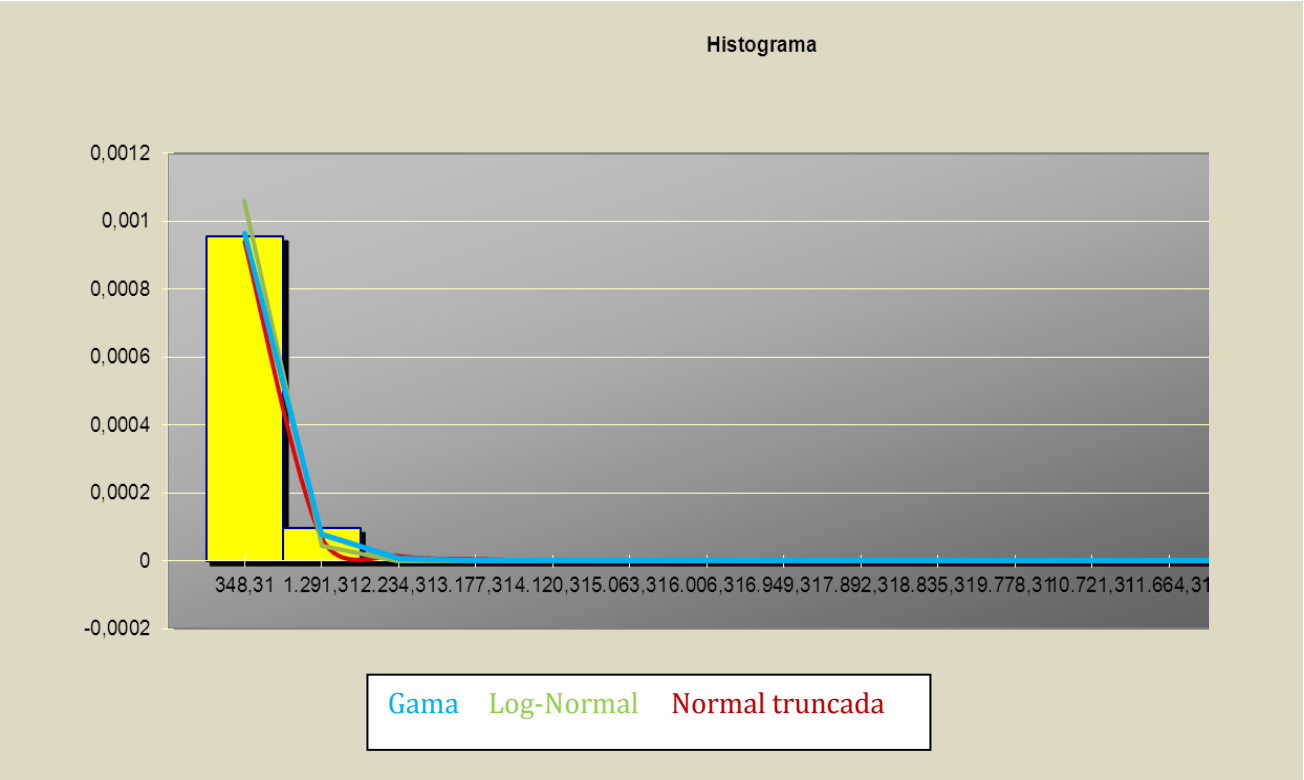



**Figura 29:** Histograma da soma dos valores apresentados agregados por Pessoa Segura - vap

Aparentemente, das distribuições selecionadas - Lognormal, Gamma e Normal truncada - a que mais se ajusta à amostra é a Gamma.

Fizeram-se testes de ajustamento à distribuição, cuja estatística de teste se encontra no Anexo 1, utilizando o teste do Qui-quadrado. Como se trata de uma amostra de dimensão muito grande – 242.689 clientes utilizadores – o que torna um teste muito exigente, conduzindo quase certamente à rejeição de qualquer distribuição, houve necessidade de fazer subamostras.

Construíram-se, então, 1000 subamostras de menor dimensão e registámos a partir dessas novas amostras a quantidade de sucessos, definindo como sucesso a "não rejeição" da distribuição em teste. Desta forma ultrapassou-se o problema, sem desprezar a quantidade de informação.

Para cada dimensão de amostra abaixo descrita, obtiveram-se os seguintes resultados:

| | | Nível de Significância | | |
|---|---|---|---|---|
| Dimensão das Subamostras | | 10% | 5% | 1% |
| n=10% Amostra | 24.269 | 17,0% | 24,8% | 40,7% |
| n=1% Amostra | 2.427 | 54,4% | 64,3% | 85,3% |
| n=0,1% Amostra | 243 | 90,2% | 90,6% | 90,8% |

**Tabela 8:** % de subamostras com "não rejeição" da distribuição Gamma

Assim, concluiu-se pela adequabilidade da distribuição a utilizar na definição do modelo linear generalizado porque se optou.

Estes resultados foram confirmados, para as amostras de menor dimensão, com o teste de *Kolmogorov-Smirnov*.

#### 3.2.2.2 Ensaios de GLM – Modelo Gamma

Neste processo foram ensaiados dois tipos de modelos os lineares com variável transformada e os lineares generalizados.

Nos modelos lineares de variável transformada, e apesar de, pela distribuição ajustada à variável resposta, a transformação logarítmica parecer muito mais indicada, ensaiámos também a transformação da raíz-quadrada, que tantas vezes tem sucesso, embora, como se pode ver pelas figuras abaixo, mantenha a sua distribuição muito enviesada.

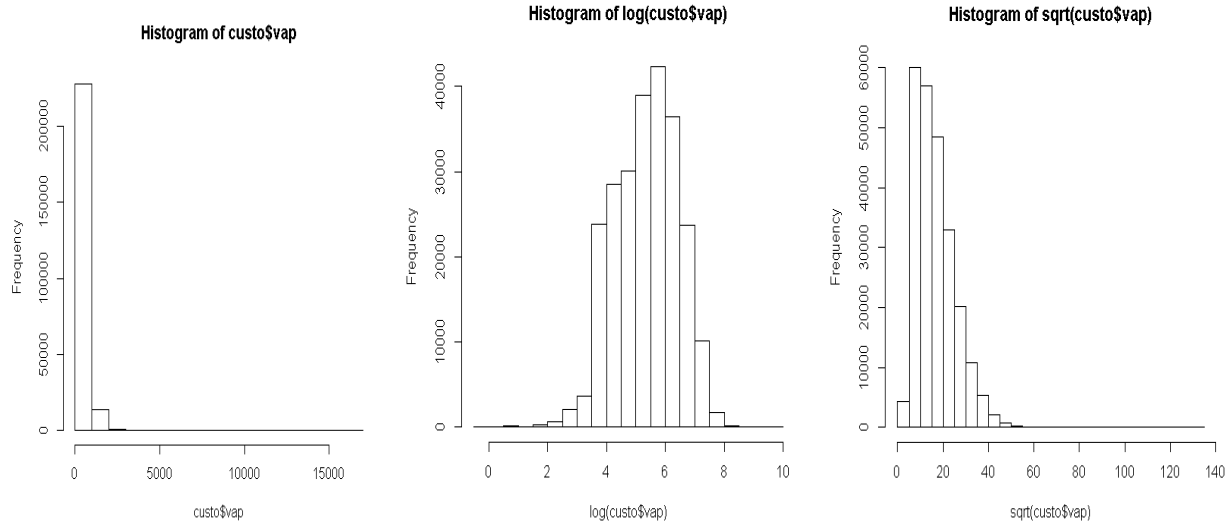



**Figura 30:** Representação dos resíduos com as várias aproximações à variável endógena

Optou-se por fazer a modelação usando os modelos lineares generalizados para todas as hipóteses porque, desta forma, a estimativa dos $\hat{\beta}$'s é determinada através dos estimadores de máxima verosimilhança para todas as distribuições, além de que admite diferentes variâncias para os erros aleatórios $\varepsilon$'s. Quando os resíduos são normais os parâmetros estimados coincidem com as estimativas do método dos mínimos quadrados.

Com suporte no software R foram então ensaiados 3 modelos que convergiram: 2 modelos lineares com função link logarítmica e raiz-quadrática e o último com base na função Gamma e função link logarítmica. A insistência nos modelos gaussianos prende-se com a literatura consultada que, por vezes, acaba por registar melhores resultados nestes[15].

Antes de entrar nos resultados dos modelos, vamos, tal como já foi referido, começar por selecionar os *clusters* que vamos adoptar para a análise de regressão.

Apresentamos de seguida os resultados obtidos com os três modelos selecionados para cada um dos conjuntos de *clusters*:

[15] *Exemplo encontrado no Artg.º "Modelling Individual Patient Hospital Expenditure for General Practice Budgets",* Hugh Gravelle e outros,Centre for Health Economics, The University of York

| Tot Y | 84.531.147 | | | |
|---|---|---|---|---|
| **Distrito** | **TotY^** | **RMSE** | **MAPE** | **$R^2$** |
| Clusters Hierq | | | | |
| Lm Ln(Y) | 84.547.615 | 362,2809 | 251,7993 | 7,32% |
| LM sqrt(Y) | 84.530.981 | 362,7273 | 252,501 | 7,09% |
| Gamma(Ln) | 84.400.318 | 361,8403 | 251,4168 | 7,55% |
| Clusters Kmeans | | | | |
| Lm Ln(Y) | 84.551.947 | 362,5189 | 251,9639 | 7,20% |
| LM sqrt(Y) | 84.531.050 | 362,934 | 252,6449 | 6,99% |
| Gamma(Ln) | 84.411.236 | 362,1105 | 251,6186 | 7,41% |

| **ZonaM** | **TotY^** | **RMSE** | **MAPE** | **$R^2$** |
|---|---|---|---|---|
| Clusters Hierq | | | | |
| Lm Ln(Y) | 84.518.978 | 362,1393 | 251,7637 | 7,40% |
| LM sqrt(Y) | 84.531.147 | 362,6924 | 252,558 | 7,11% |
| Gamma(Ln) | 84.375.101 | 361,7179 | 251,392 | 7,61% |
| Clusters Kmeans | | | | |
| Lm Ln(Y) | 84.542.287 | 362,0329 | 251,7437 | 7,45% |
| LM sqrt(Y) | 84.531.058 | 362,6171 | 252,7175 | 7,15% |
| Gamma(Ln) | 84.363.308 | 361,6371 | 251,3459 | 7,65% |

**Tabela 9:** Resultados da Regressão Linear Generalizada para as várias classificações elaboradas

A classificação que conduz a melhores resultados é diferente consoante se trata de Distrito ou Zona Multicare. Não há, no entanto, grande diferença entre todas elas. Assim, vamos optar pelas variáveis explicativas "Distritos", pelas razões explanadas anteriormente (pág.48), e pela classificação hierárquica.

Uma vez que se procura encontrar variáveis explicativas para a tarifa, e portanto, mais do que se fazer previsão individual de custos, se pretender explicar comportamentos médios previsíveis, optou-se por representar as médias das estimativas obtidas pelo modelos - ver abaixo.

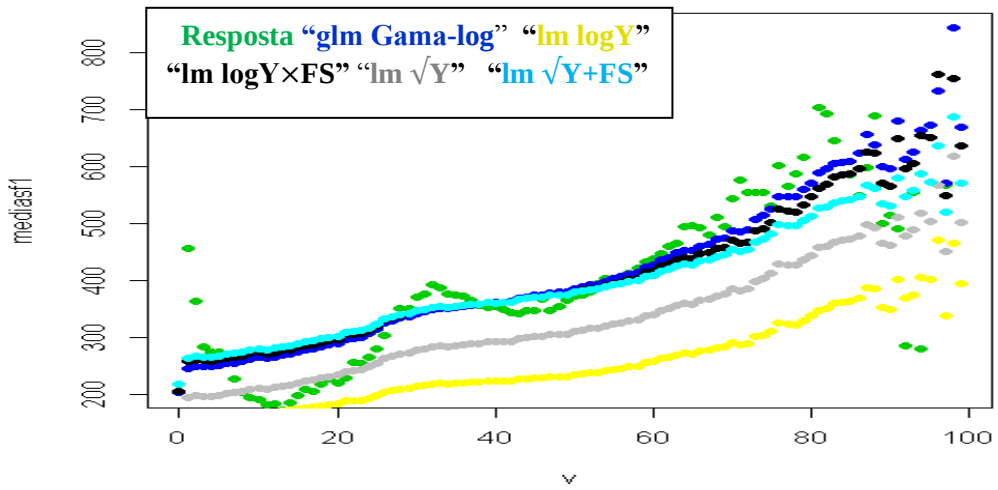



**Figura 31:** Médias por idade das estimativas obtidas pelos modelos de regressão com e sem correção do fator Smearing

Conforme se pode observar, é notório o efeito da correção do factor *smearing* – passagem da curva em amarelo para a negra e da curva em cinzento para azul turquesa – nas estimativas que obtivémos pelos modelos lineares de variável endógena transformada.

Ficamos então com os três modelos[16] "glm Gama-log", "lm LogY ×FS" – linear com a transformação logarítmica corrigida pelo fator *smearing* – e o "lm √Y + FS" – linear com a

[16] *Serão sempre estes três modelos a ser ensaiados pelo que, em todas as referências futuras, as estimativas através da regressão linear de variável dependente transformada apresentam-se já corrigidas pelo factor smearing, mesmo que não se faça referência a esse facto.*

transformação da raiz quadrada corrigida pelo fator *smearing* – que apresentam estimativas relativamente próximos da variável resposta. No entanto não acompanham o perfil de custos sobretudo nas idades mais jovens (até aos cinquenta anos).

De facto, esta suspeição - de que um modelo único poderia não ter qualquer adesão aos valores observados, mesmo em termos médios - tinha sido levantada quando se fizeram as análises preliminares das variáveis explicativas (Idade, Sexo e Parentesco). Fez-se, por isso, uma remodelação de acordo com as seguintes faixas etárias:

1. Infância – menores de 13 anos
2. Juventude – dos 14 aos 49 anos
3. Maturidade – maiores de 50 anos

### Infância

Com uma amostra de quarenta e oito (48.269) mil indivíduos ensaiaram-se os seguintes modelos:

| Infância | $\sum \widehat{Vap}$ | RMSE | MAPE | $R^2$ |
|---|---|---|---|---|
| **GLM - Gamma** | 12.429.152 € | 236,9086 | 164,1113 | 9,712% |
| **OLS - Ln(Y)** | 12.587.178 € | 236,9108 | 164,5286 | 9,709% |
| **OLS - $\sqrt{Y}$** | 12.451.254 € | 237,4385 | 164,7517 | 9,307% |

**Tabela 10:** Avaliação dos resultados dos modelos para a Infância

Todos os modelos apresentam medidas de avaliação muito próximas e relativamente pouco significativas. O Modelo Gamma, apesar de tudo, é o que apresenta melhores indicadores, embora a soma dos valores estimados esteja abaixo dos valores observados ($\sum_{i=1}^{n_1} \hat{Vap}$ =12.451.254€).

### Juventude

Nesta nova faixa etária encontram-se a maioria das pessoas seguras ficando com uma amostra de cento e quarenta (141.359) mil indivíduos.

Um dos eventos da vida que conduz a uma maior utilização do seguro de saúde é o nascimento dos filhos. Para levar em consideração este facto acrescentámos como variável explicativa deste modelo uma variável construída a partir das taxas de fecundidade apresentadas nas Estatísticas Demográficas do INE para 2010, tal como referido no **Anexo 2**, conseguindo assim um ajustamento de melhor qualidade, em termos de valores médios como se pode observar no gráfico abaixo:

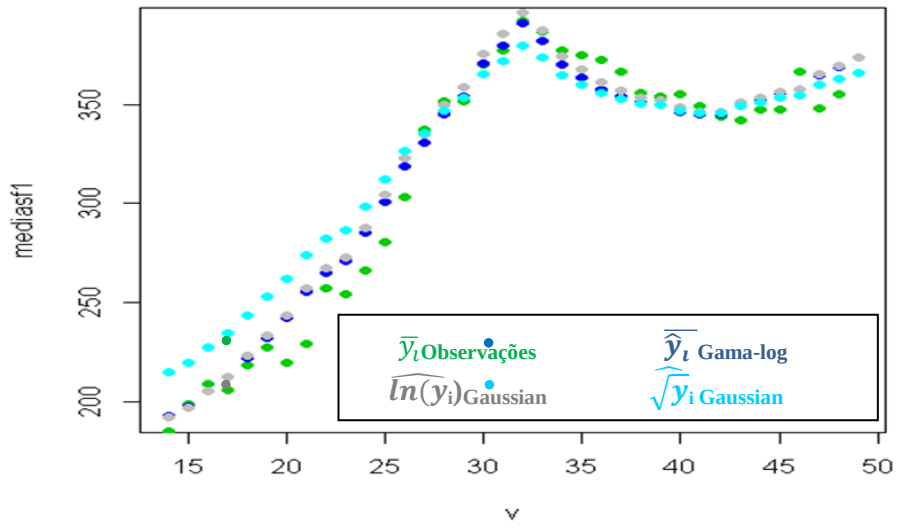



**Figura 32:** Médias por idade das estimativas obtidas pelos modelos de regressão para a Juventude

Se calcularmos os valores médios por género também as estimativas se aproximam bastante dos valores médios observados, conforme se pode observar na Figura abaixo:

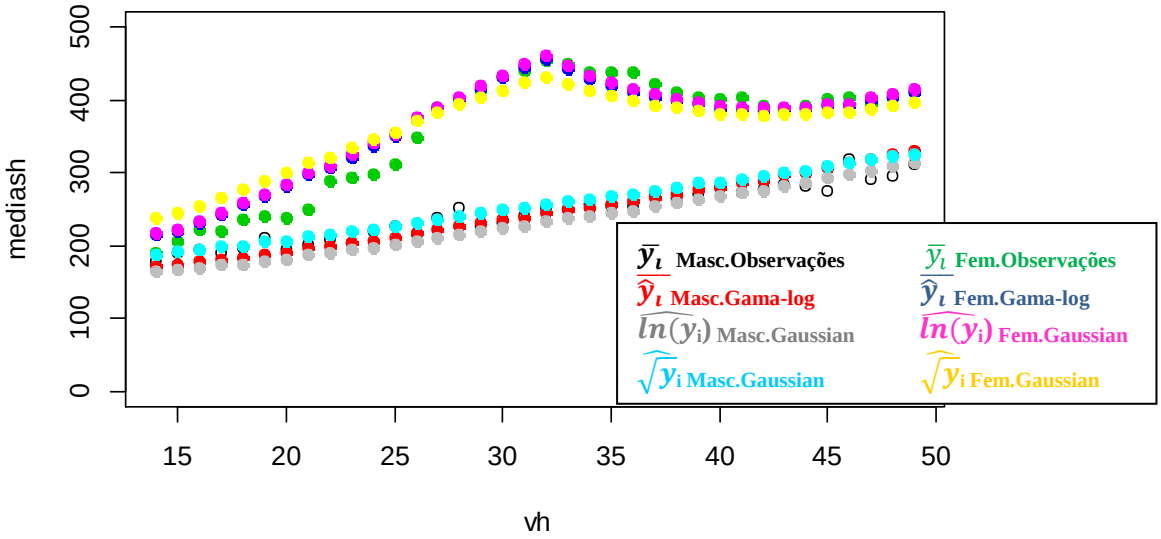



**Figura 33:** Médias por idade e por género das estimativas obtidas pelos modelos de regressão para a Juventude

Os resultados dos modelos ensaiados foram os seguintes:

| Juventude | $\widehat{\sum Vap}$ | RMSE | MAPE | $R^2$ |
|---|---|---|---|---|
| GLM – Gamma | 48.319.336 € | 349,6327 | 247.8574 | 7,959% |
| OLS – Ln(Y) | 48.715.306 € | 349,8161 | 248,5106 | 7,863% |
| OLS - $\sqrt{Y}$ | 48.376.460 € | 350,3397 | 233,4312 | 7,587% |

**Tabela 11:** Avaliação dos resultados dos modelos para a Juventude

De novo os modelos apresentam medidas de avaliação muito próximas e relativamente pouco significativas. O Modelo Gamma continua a ser o que apresenta melhores indicadores, embora a soma dos valores estimados esteja abaixo dos valores observados ($\sum_{i=1}^{n_1} \hat{Vap}$ =48.376.460€).

## Maturidade

Esta última faixa etária de cinquenta e três (53.060) mil indivíduos apresenta uma grande dispersão de valores, sobretudo nas últimas idades onde existe um número relativamente baixo de indivíduos.

Optou-se por isso por fazer duas regressões uma com todos os indivíduos e outra apenas com os indivíduos de idades compreendidas entre os 50 e os 70 anos, que designamos por modelo da Maturidade truncada, e que corresponde a uma amostra de quarenta e sete (46.546) mil indivíduos.

E o resultado dos modelos ensaiados foram os seguintes:

| Maturidade | $\sum Vap$ | RMSE | MAPE | $R^2$ |
|---|---|---|---|---|
| **GLM - Gamma** | 23.725.738 € | 456,9757 | 317.7498 | 5,147% |
| **OLS - Ln(Y)** | 23.915.452 € | 457,9898 | 318,7328 | 4,725% |
| **OLS - $\sqrt{Y}$** | 23.703.434 € | 457,0902 | 317,8586 | 5,099% |

**Tabela 12:** Avaliação dos resultados dos modelos para a Maturidade

Tal como era expectável, por se tratar de uma faixa etária com grande dispersão de valores, o $R^2$ apresenta valores ainda mais baixos do que nos estudos anteriores ($\cong 5\%$) O Modelo Gamma continua a ser o que apresenta melhores indicadores, embora a soma dos valores estimados esteja abaixo dos valores observados ( $\sum_{i=1}^{n_2} \widehat{Vap}_i = 23.703.434€$).

Para o Modelo da Maturidade Truncada, eis os resultados obtidos:

| Maturidade Truncada | $\sum \widehat{Vap}$ | RMSE | MAPE | $R^2$ |
|---|---|---|---|---|
| **GLM - Gamma** | 19.939.902 € | 423,2540 | 304.3815 | 4,536% |
| **OLS - Ln(Y)** | 20.145.890 € | 424,2133 | 305,5535 | 4,103% |
| **OLS - $\sqrt{Y}$** | 19.927.985 € | 423,3234 | 304,5834 | 4,505% |

**Tabela 13:** Avaliação dos resultados dos modelos para a Maturidade abaixo dos 70 anos

Como se pode observar a capacidade explicativa do modelo não melhora, o $R^2$ até baixa, e o erros médios quadráticos e absolutos apenas reduzem cerca de 7,4% e 4,2% respectivamente.

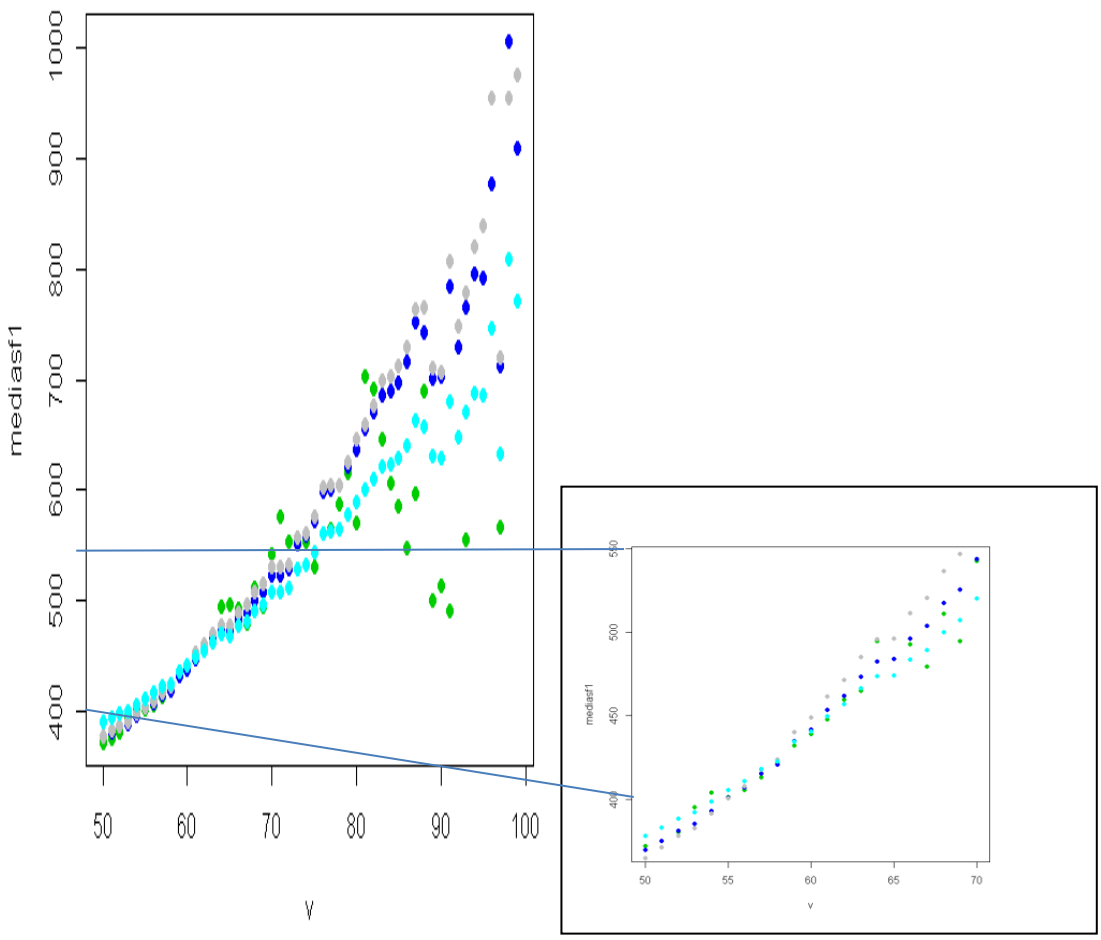



**Figura 34:** Comparação das médias, por idade, das estimativas obtidas para a Maturidade e Maturidade truncada

Embora as estimativas obtidas pelos modelos considerados sejam individuais, o objectivo último é explicar a variabilidade dos comportamentos médios idade a idade. Assim, além das estimativas devolvidas automaticamente pelo software, foi necessário calcular, de forma adicional, manualmente, as estimativas que o modelo truncado, produziria nos indivíduos com mais de 70 anos. Esta informação encontra-se resumida na Figura 35, onde se apresentam as médias das estimativas para os dois modelos:

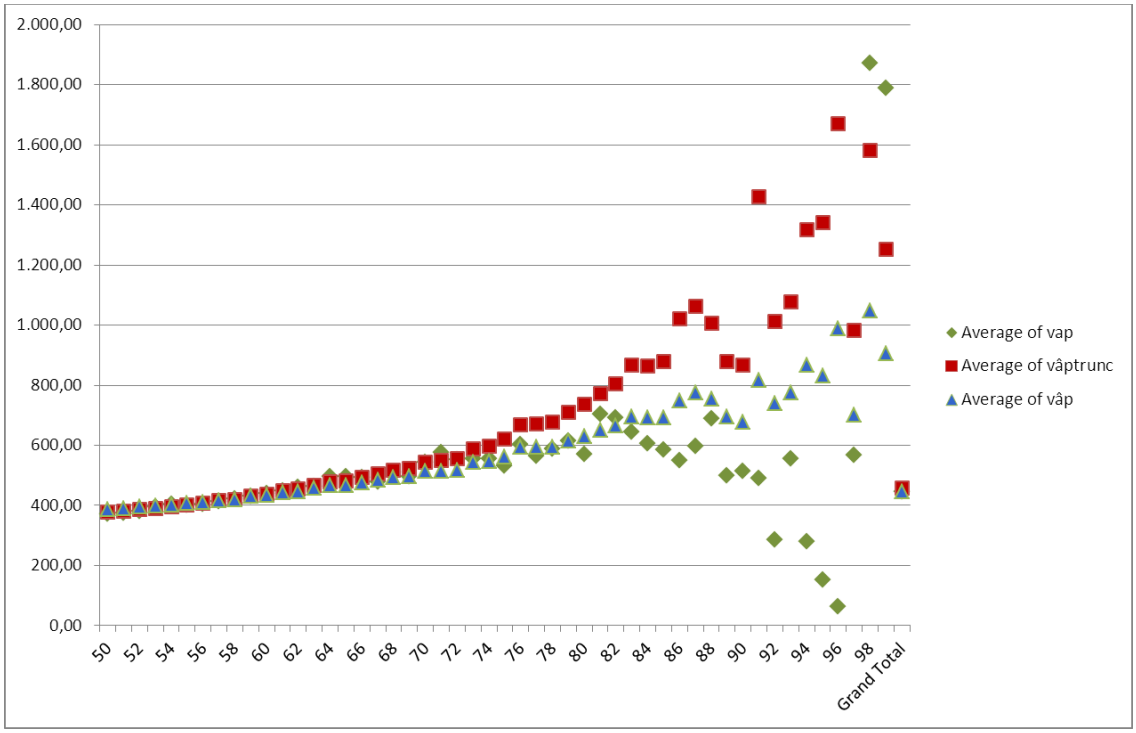



**Figura 35:** Médias por idade das estimativas obtidas pelos modelos de regressão para a Maturidade com projeção do modelo truncado para as idades acima dos 70 anos

De côr azul encontram-se as médias dos valores apresentados pelos indivíduos com mais de 50 anos. As estimativas para estes indivíduos a partir do modelo truncado (vermelho) acabam por apresentar valores médios muito acima dos valores observados na nossa amostra, que, nesta idades mais avançadas e como sabemos, apresenta uma dimensão reduzida.

## Modelo Global

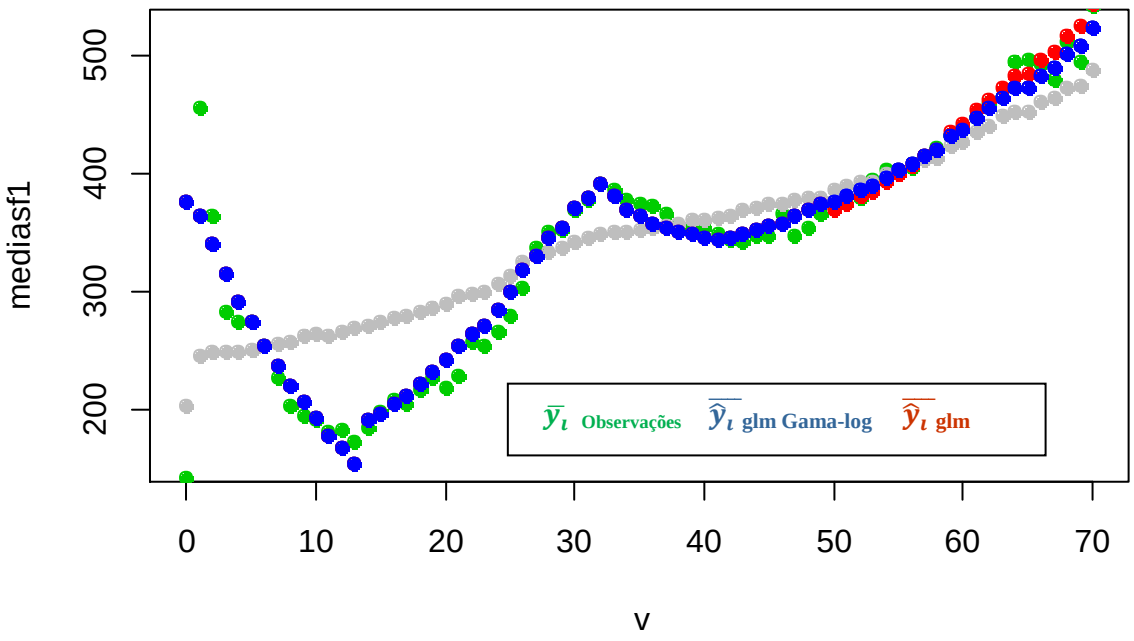



**Figura 36:** Representação das médias, por idade, das estimativas de custo para a junção dos modelos: Infância, Juventude e Maturidade

Quando juntamos os vários modelos, mas olhando agora para as idades inferiores aos 70 anos, classe onde se posicionam 97% dos clientes utilizadores e 98% das Pessoas Seguras na cobertura de ambulatório, verifica-se o que segue:

| Modelo Global de Custo | $\sum \widehat{Vap}$ | RMSE | MAPE | $R^2$ |
|---|---|---|---|---|
| **Total** | 80.258.615 € | 346,3209 | 241.8187 | 9,198% |
| **Truncado** | 80.688.390 € | 346,3294 | 241,8814 | 9,194% |

**Tabela 14:** Avaliação dos resultados para a junção dos modelos: Infância, Juventude e Maturidade

Aparentemente estes dois modelos apresentam uma qualidade muito próxima no que respeita a variabilidade explicada da amostra, sendo que o modelo não truncado, e ambos limitados à amostra das Pessoas Seguras até aos 70 anos de idade, explica mais variabilidade.

A variância desta amostra truncada é 132.087,7. O erro quadrático médio (RMSE) do primeiro modelo é 346,3209 enquanto o do modelo truncado346,3294. Ou seja, o modelo truncado tem o maior RMSE logo menor $R^2$. Ainda podemos, complementarmente, olhar para uma medida de verosimilhança desta amostra, com base em cada um destes modelos. Como estes modelos têm o mesmo número de β's, podemos fazê-lo através da *Deviance*.

### 3.2.3 Modelo de Ocorrência

Resta-nos então construir o modelo de ocorrência que se enraizará em todas as pessoas seguras e não exclusivamente apenas nos clientes utilizadores como aconteceu com os modelos de custo. A nova amostra tem trezentas e oitenta e três mil (382.947) Pessoas seguras.

Esta variável tem indiscutivelmente uma distribuição de *Bernoulli*. Assim os modelos mais adequados seriam o Logit ou o Probit. Optou-se, conforme será referido na próxima secção, pelo Modelo Logit, definido em 2.2.1.3.

#### 3.2.2.3 Ensaios de GLM – Modelo Logit

Foram utilizadas, nesta regressão, as mesmas variáveis dummy e as mesmas partições da idade que foram utilizadas para o Modelo de Custo. Não obstante, ensaiou-se um modelo único (para todas as idades), onde se voltou observar que o modelo, em termos de média das estimativas, ficava muito desajustado da realidade:

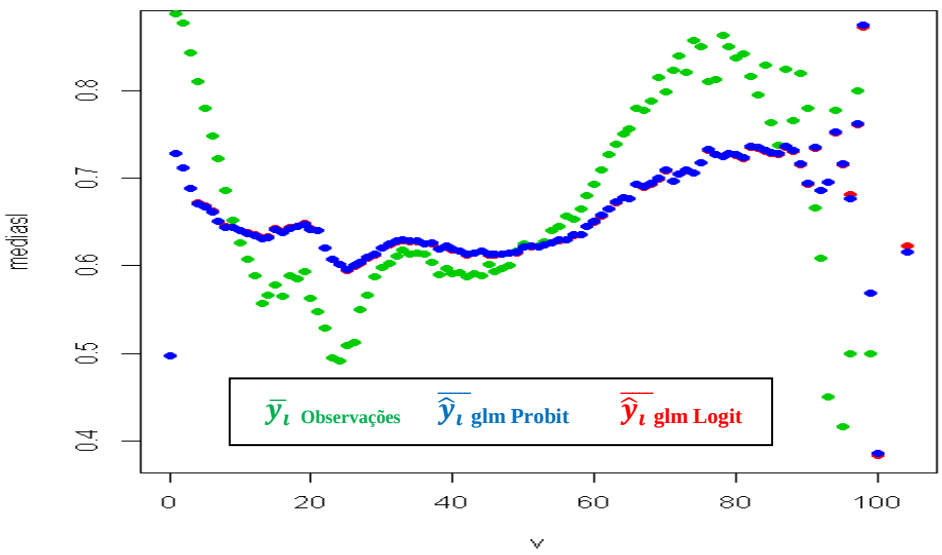



**Figura 37:** Representação das médias, por idade, das estimativas de ocorrência

Assim se confirma a necessidade da partição das idades.

## Infância

Nesta amostra encontram-se sessenta e seis mil (66.233) pessoas seguras e a representação das médias das estimativas aparecem muito mais alinhadas, do que as médias das estimativas para as mesmas idades no modelo global.

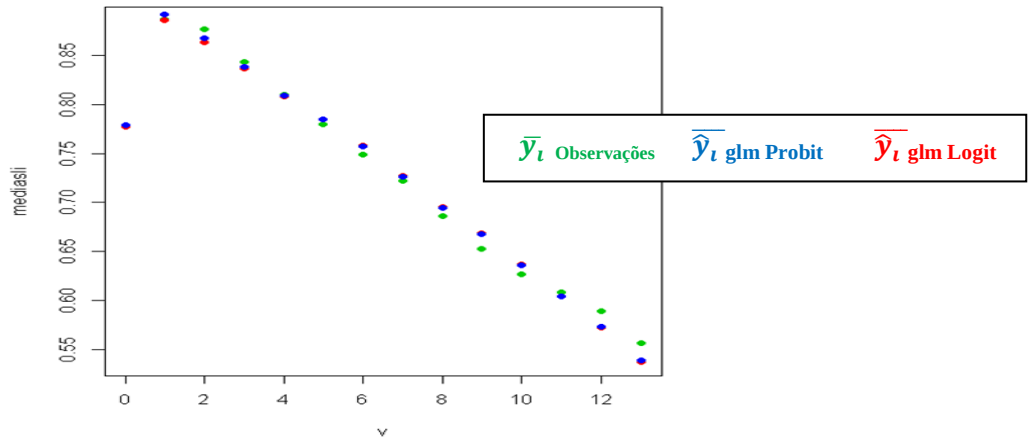



**Figura 38:** Representação das médias, por idade, das estimativas de ocorrência para o modelo da Infância

| Infância | $\sum \widehat{Ocorr}$ | RMSE | MAPE | $R^2$ | AUC |
|---|---|---|---|---|---|
| Logit | 48.283 | 0,4193 | 0,3502 | 11,01% | 71,85% |
| Probit | 48.329 | 0,4191 | 0,3495 | 11,11% | 71,97% |

**Tabela 15:** Avaliação dos resultados para o modelo de Infância

Observa-se uma melhoria muito ligeira quando se opta pelo modelo Probit em deterimento do primeiro modelo, no entanto, e porque este usa a função ligação canónica, vamos optar pelo modelo Logit que apresenta melhores estimadores do que o Probit[17].

## Juventude

Esta segunda amostra, representada na Figura 39 , é constituída pelos indivíduos com idades entre os 14 e os 49 anos:

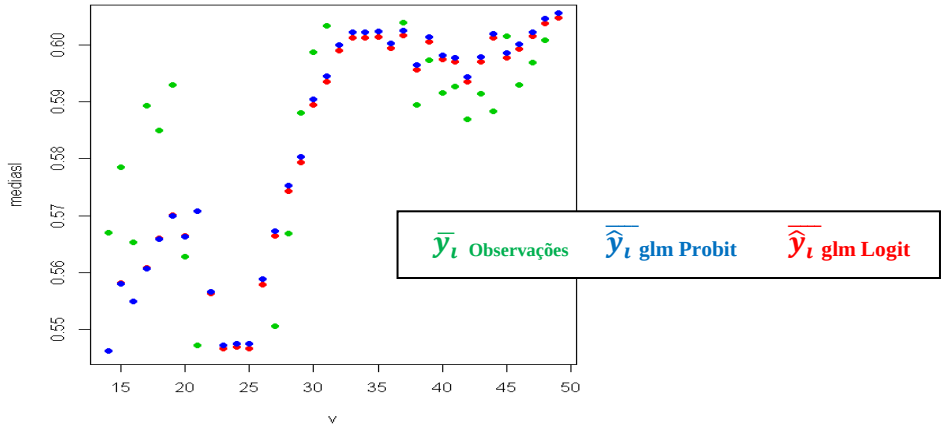



**Figura 39:** Representação das médias, por idade, das estimativas de ocorrência para o modelo da Juventude

| Juventude | $\sum \widehat{Ocorr}$ | RMSE | MAPE | $R^2$ | AUC |
|---|---|---|---|---|---|
| Logit | 141.434 | 0,4542 | 0,4122 | 14,77% | 72,71% |
| Probit | 141.621 | 0,4541 | 0,4121 | 14,80% | 72,72% |

**Tabela 16:** Avaliação dos resultados para o modelo de Juventude

Os resultados deste modelo são muito semelhantes aos dos modelos de infância e, portanto, as conclusões são as mesmas. Daí se optar pelo modelo Logit.

## Maturidade

Em analogia com o modelo de Custo foram ensaiados os modelos para a amostra dos indivíduos com 50 anos ou mais, como ilustrado na Figura XXX1. Da mesma forma, na Figura XXX2, representámos o modelo truncado com as pessoas seguras até aos 70 anos.

[17] Amaral Turkman, M. A. e Silva, G. (2000). Páginas 14 e 15

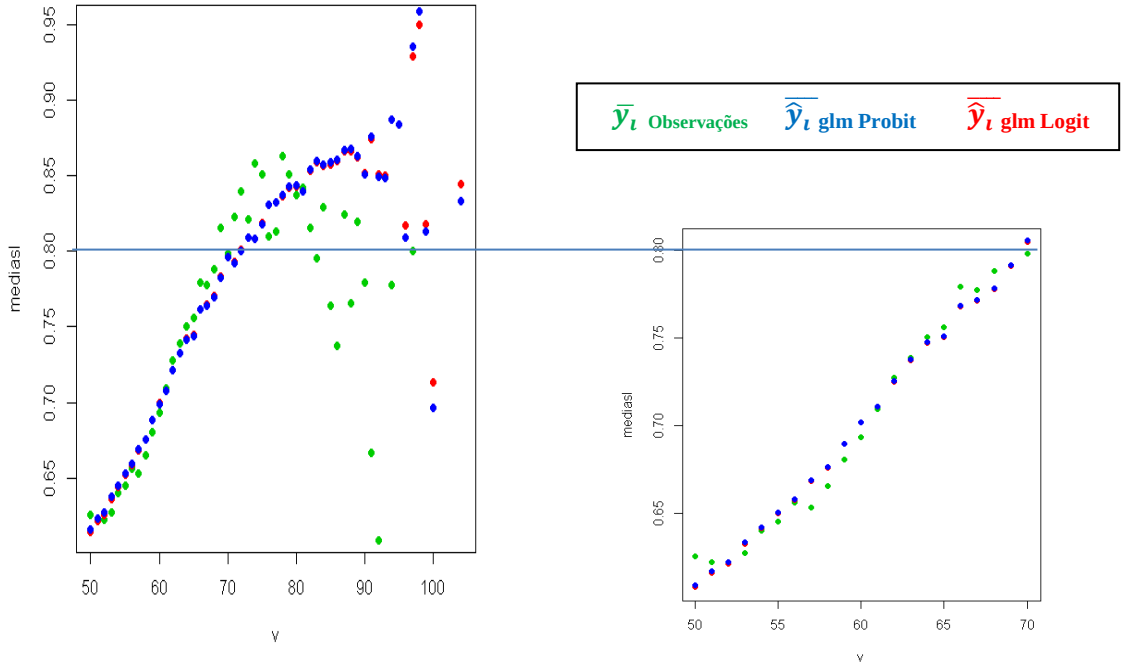



**Figura 40:** Comparação das médias, por idade, das estimativas de ocorrência obtidas para a Maturidade e Maturidade truncada

| Maturidade | $\sum \widehat{Ocorr}$ | RMSE | MAPE | $R^2$ | AUC |
|---|---|---|---|---|---|
| Logit | 53.081 | 0,4237 | 0,3585 | 15,64% | 74,33% |
| Probit | 53.118,22 | 0,4236 | 0,3584 | 15,69% | 74,34% |

**Tabela 17:** Avaliação dos resultados para o modelo de ocorrência da Maturidade

| Maturidade Truncada | $\sum \widehat{Ocorr}$ | RMSE | MAPE | $R^2$ | AUC |
|---|---|---|---|---|---|
| Logit | 46.563 | 0,4305 | 0,3702 | 15,21% | 73,74% |
| Probit | 46.598,87 | 0,4304 | 0,3700 | 15,21% | 73,75% |

**Tabela 18:** Avaliação dos resultados para o modelo de ocorrência da Maturidade truncada

Tal como nos modelos ensaiados nas restantes faixas etárias, observa-se que o modelo Probit não traz nenhum benefício significativo aos dados em estudo, para além das desvantagens teóricas já mencionadas.

No que respeita à selecção entre os dois modelos de maturidade, vamos avaliá-los sob os mesmos dados: entre os 50 e os 70 anos, onde incide a maior parte dos indivíduos seguros em idade madura e verificar se, pelo menos nesse domínio apresenta maior nível de concordâncias.

Assim a tabela seguinte apresenta os resultados do modelo Logit para a Maturidade e a Maturidade Truncada, mas restringido aos indivíduos com idade entre os 50 e os 70 anos:

| Modelo Logit | $\sum \widehat{Ocorr}$ | RMSE | MAPE | $R^2$ | AUC |
|---|---|---|---|---|---|
| **Maturidade** | 46.570 | 0,4306 | 0,3700 | 15,13% | 73,76% |
| **Maturidade Truncada** | 46.563 | 0,4305 | 0,3702 | 15,21% | 73,74% |

**Tabela 19:** Avaliação dos resultados para o modelo de ocorrência Logit para a Maturidade e para a Maturidade truncada

Não se observando melhoria significativa, nem mesmo quando se limita as análises às idades dos 50 aos 70 anos, opta-se pelo Modelo não truncado para a constituição do modelo global.

### Modelo Global

Juntando, para as três faixas etárias, os modelos selecionados anteriormente e comparando-os com o modelo global inicial, verifica-se, como se pode observar na Figura 41, uma clara melhoria no ajustamento das médias das estimativas por idade às médias das observações.

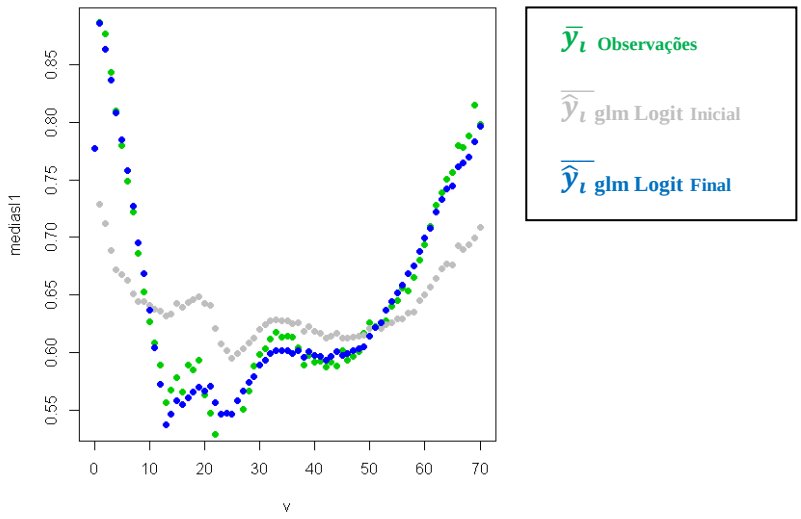



**Figura 41:** Representação das médias, por idade, das estimativas do modelo Logit de ocorrência para a junção dos modelos: Infância, Juventude e Maturidade

Desta forma, sempre que nas próximas secções se mencionar o modelo de ocorrência, este refere-se ao modelo global quando juntamos os três modelos logit para as respectivas faixas etárias.

## 3.2.4 Modelo Combinado

O Modelo combinado será o modelo resultante da combinação do modelo global de ocorrência com o modelo global de custo para a estimativa do custo do risco, reconhecido na teoria do risco como modelo das indemnizações agregadas.

Na secção anterior definimos qual seria o modelo global de ocorrência a utilizar. Já no que respeita ao modelo global de custo, não era claro qual o modelo que produzia melhor ajustamento no que respeita à utilização, ou não, da amostra truncada no modelo da maturidade. Face à opção

feita para o modelo de ocorrência vamos optar também pelo modelo de custo não truncado por uma questão de coerência de opções.

Estão finalmente reunidas as condições para construír o modelo combinado com base nos desenvolvimentos apresentados na secção 2.2.5.

$$Y_i = Z_i \times U_i \text{ , com i=1,…,382.947}$$

No que respeita aos primeiros momentos (centrado de 2ªordem e não centrados) tem-se que:

- $E[Y_i] = E[Z_i] \times E[U_i]$
- $Var[Y_i] = Var[Z_i] \times E[U_i] + E^2[Z_i] \times Var[U_i]$

Através das medidas que definimos, e que temos utilizado ao longo deste estudo, podemos avaliar a qualidade do ajustamento, concluindo-se que conseguimos explicar 18% da variabilidade da amostra.

| Maturidade | $\sum y_i$ | $\sum \hat{y}_i$ | RMSE | MAPE | $R^2$ |
|---|---|---|---|---|---|
| Logit . Gama | 84.531.147,39€ | 84.976.078,72€ | 310,79 | 201,32 | 18,08% |

**Tabela 20:** Avaliação dos resultados para o modelo de ocorrência Logit para o modelo combinado (Logit . Gamma)

A construção da tarifa suporta-se na média de cada classe tarifária que for definida a partir das variáveis explicativas selecionadas.

Com base na actual tarifação, vamos supor que estas classes serão definidas apenas com a idade do indivíduo, o que leva à necessidade de conhecer os primeiros momentos da média dos indivíduos:

$$\text{Custo do Risco}_{\text{idade}} = \text{E}[\text{X}_{\text{idade}}] + \Phi^{-1}(0,9) \text{ x } \sqrt{Var[X_{idade}]/n_{idade}}$$

Na figura 42 apresentamos as estimativas de custo do risco por idade, bem como os intervalos de confiança a 90%. De igual forma, no Anexo 3: apresentamos as estimativas a 70% e a 80% de confiança.

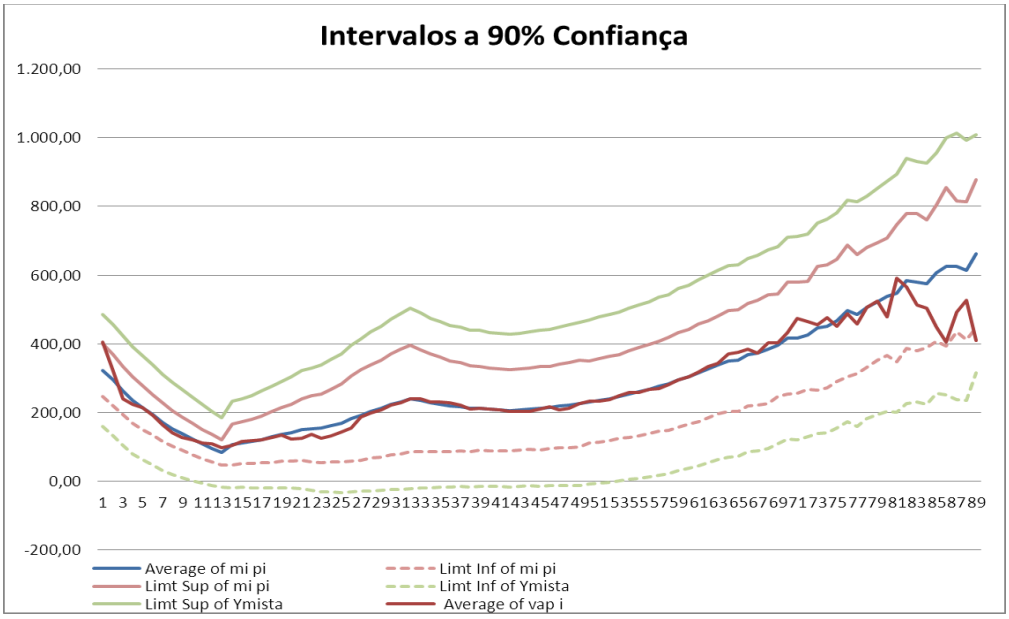



**Figura 42:** Representação dos resultados médios e intervalos de confiança para cada idade no modelo Global

Conforme se pode observar, as médias das estimativas ( Average of "mipi") estão muito próximas das médias dos valores observados (Average of "vap i") à excepção dos últimos anos, onde a pouca população das classes promove grande oscilação(variabilidade) dos resultados.

A amplitude dos intervalos de confiança - [Limt Inf of Ymista, Limt Sup of Ymista] - obtidos com base na aproximação à Normal, mas considerando a estrutura do modelo global ($Y_i = Z_i \times U_i$) , é naturalmente superior, mas com uma margem de erro inferior por se tratar de uma estatística mais exacta.

O segundo intervalo- [Limt Inf of mipi, Limt Sup of mipi] – que é construído apenas com a média e o desvio padrão amostrais, apresenta uma menor amplitude uma vez que se suporta na amostra que serviu de suporte à Modelização Linear Generalizada, mas originará naturalmente uma maior margem de erro na tarifação.

A título de exemplo apresentamos, para um conjunto de idades, quatro dos custos de risco resultantes, se suportados pelo percentil de 90%:

| Idades | C. Risco |
|---|---|
| 15 | 164,36 |
| 35 | 322,16 |
| 55 | 364,48 |
| 75 | 596,38 |

**Tabela 21:** Custos de Risco com base no percentil 90%

# 4. Conclusões

O objectivo primordial deste trabalho era identificar, dentro das variáveis constantes na base de dados da Seguradora sobre os riscos em carteira, quais as que poderiam ser utilizadas para valorizar o risco e explicar a sua variabilidade. Foram selecionadas carterísticas de duas dimensões: Pessoas Seguras e produtos em vigor.

A primeira conclusão é que, a partir da informação constante nestes dados, se consegue uma valorização do risco com um desvio de +0,5%([18]) mas apenas conseguimos explicar a variabilidade em 18,1%; isto é, serve para estimar comportamentos médios previsíveis, da grande utilidade na construção de uma tarifa, mas não serve para estimar comportamentos individuais.

O modelo que se construiu para explicar a utilização do seguro – o modelo Logit – atingiu um nível de concordâncias de 74%, o que o classifica como modelo com capacidade de discriminação aceitável.

No que respeita ao modelo de custo, por se ter subdividido em três modelos por faixas etárias (infância, juventude e maturidade) conseguimos apenas passar de um modelo com $R^2$ de 7% para um modelo  9%, justificado, sobretudo, pela grande variabilidade das observações acima dos 50 anos.

Finalmente, e falando agora das variáveis explicativas, quase todas tem um pequeno poder explicativo (todos $\beta$ inferiores a um, à excepção do $\beta_0$ e do coeficiente da variável "taxa de fecundidade" utilizada no modelo de custo da juventude). No entanto, e para as variáveis de baixo coeficiente, é de referir que:

- A idade, o sexo, o tipo de seguro (individual ou grupo) e o número de coberturas do produto são significativos em todos os modelos;
- O capital seguro, no modelo de custo, só se revelou significativo durante a Maturidade, em todas as idades mais jovens não se apresenta relevante. Mesmo neste modelo ocorre um coeficiente reduzido($\beta <$ 3,020e-12). Já no modelo de ocorrência, esta variável aparece sempre como variável explicativa significante, apesar de no modelo para a infância apresentar um elevado *p-value*(0,14036);
- No que respeita à percentagem de comparticipação a situação inverte-se, surge em todos os modelos de custo com significância (*p-value* <5,88e-12) e não são significativos nos modelos de ocorrência;
- Entre os parentes, os modelos dão relevância à diferença significativa entre os titulares e os cônjuges e filhos; e, apenas nos modelos de custo e no modelo de ocorrência na infância, para os outros parentes;
- No que respeita ao canal de vendas existe diferença significativa entre o canal bancário, as redes tradicionais e as restantes formas de comercialização. Apenas o modelo de Custo para a infância não atribui diferença significativa entre o canal bancário e as outras redes não tradicionais;
- Por último, a classificação dos distritos apenas é significativa para Lisboa, Porto e Setúbal e os restantes, com excepção de algumas ilhas que também se distinguiram: Porto Santo, Flores e São Jorge em quase todos os modelos.

[18] Cálculo dos desvio na valorização do risco $\sum_i \hat{Y}_i \Big/ \sum_i Y_i - 1$

Em conclusão, existe hoje, nas bases de dados das Seguradoras, um conjunto de informações sobre o Risco Seguro – Pessoa Segura no contrato – que permitiria ampliar o conhecimento que se utiliza para a apreciação do risco *a priori* , de salientar o número de coberturas do produto, o canal de vendas e a zona geográfica. Sabemos que existe, ainda, um outro conjunto de dados sobre risco e que a Seguradora ao não integrar na base de dados e apenas a armazenar por digitalização, desperdiça e que poderia permitir reconhecer melhor cada risco que tarifa.

## Próximos Passos

Algumas destas conclusões estão condicionadas pelo facto de, por falta de capacidade da máquina e por conveniência comercial, se ter recorrido à classificação das variáveis explicativas para redução do número de dummies. Assim alguns testes deverão ser efectuados para verificar se cada modelo pode ser melhorado, separando estes *clusters*, apesar da análise de *clusters* ter sido enraizada em duas variáveis endógenas que se modelizaram.

Apesar destas restrições, tal análise revelou-se útil na identificação, e na confirmação em alguns casos, das variáveis com significância na explicação do uso e do custo da cobertura de Ambulatório. Assim faz sentido estender este estudo às restantes coberturas, pelo menos ao Internamento e à Estomatologia.

Um outro aspecto que deverá ser igualmente desenvolvido é um estudo sobre a informação que se poderá adquirir junto do potencial cliente, ou que não se integra na base de dados da Seguradora, e que possa trazer mais informação sobre o risco individual de saúde, bem como a eventual utilização de histórico de sinistralidade através de outras metodologias, por exemplo Equações de Estimação Generalizadas (GEE de *Generalized Estimating Equations*)

# 5. Anexos

## Anexo 1:

Testes estatísticos usados para testar o ajustamento da distribuição Gama:

$$H_0 : X \text{ tem distribuição } Gama(\upsilon/\mu, \upsilon)$$

- Qui-Quadrado

Estatística de Teste: $X^2{}_N = \sum_{i=1}^{N} \frac{(o_i - e_i)^2}{e_i}\Big|_{H_0} \cap \chi^2_\kappa$

Região de Rejeição: $R_\alpha : X^2{}_N > \chi^2_{\kappa,1-\alpha}$

Onde:

- $\kappa = N - p - 1$ denota o número de graus de liberdade do Qui-quadrado
- $\alpha$ representa o nível de significância escolhido para o teste
- $p$ corresponde ao número de parâmetros estimados
- $N$ traduz o número de classes ($C_i$) em que são agrupados os nossos dados
- $o_i$ é o número de indivíduos da amostra observados em cada classe $C_i$
- $e_i = N \times p_i$ estima o número de indivíduos esperados em cada classe $C_i$, sendo $p_i = P\big|_{H_0}[X \in C_i]$

- Kolmogorov-Smirnov

Estatística de Teste: $D = \sup_{x \in \Re_0^+} \left| F_0(x) - F_n^*(x) \right|$

Onde $F_0(x)$ denota a função de distribuição do modelo postulado em $H_0$ e $F_n^*(x)$ denota a função de distribuição empírica. Os pontos críticos e os respectivos níveis encontram-se tabelados.

## Anexo 2:

*Construção de uma variável explicativa a partir da taxa de fecundidade disponibilizada pelo INE nas Estatísticas Demográficas de 2010:*

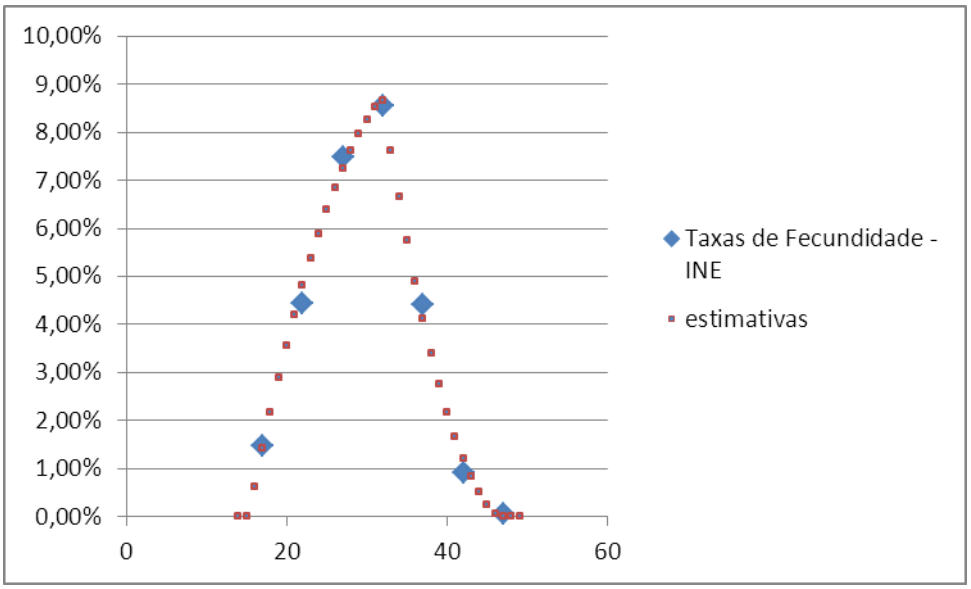
**Taxas de fecundidade por grupo etário das mulheres**

(em permilagem)

| | 2005 | 2006 | 2007 | 2008 | 2009 | 2010 |
|---|---|---|---|---|---|---|
| 15-19 | 19,00 | 17,00 | 16,90 | 16,20 | 15,50 | 14,70 |
| 20-24 | 47,60 | 45,50 | 44,10 | 45,90 | 43,80 | 44,40 |
| 25-29 | 84,30 | 79,60 | 76,10 | 76,70 | 72,70 | 74,70 |
| 30-34 | 85,30 | 83,80 | 82,80 | 85,80 | 82,50 | 85,60 |
| 35-39 | 37,60 | 38,40 | 39,40 | 42,00 | 41,60 | 44,20 |
| 40-44 | 7,40 | 7,70 | 7,40 | 7,80 | 8,00 | 9,10 |
| 45-49 | 0,50 | 0,40 | 0,30 | 0,40 | 0,50 | 0,50 |

*Esta variável auxiliar vai ser definida positiva para os indivíduos do género feminino e nula para os restantes.*

*Assumiram-se assim, para as mulheres, que as taxas apresentadas eram válidas para a idade média de cada intervalo e optou-se por fazer um ajustamento através da regressão polinomial*




*Entre os 15 e os 32 anos ajustou-se o polinómio:*

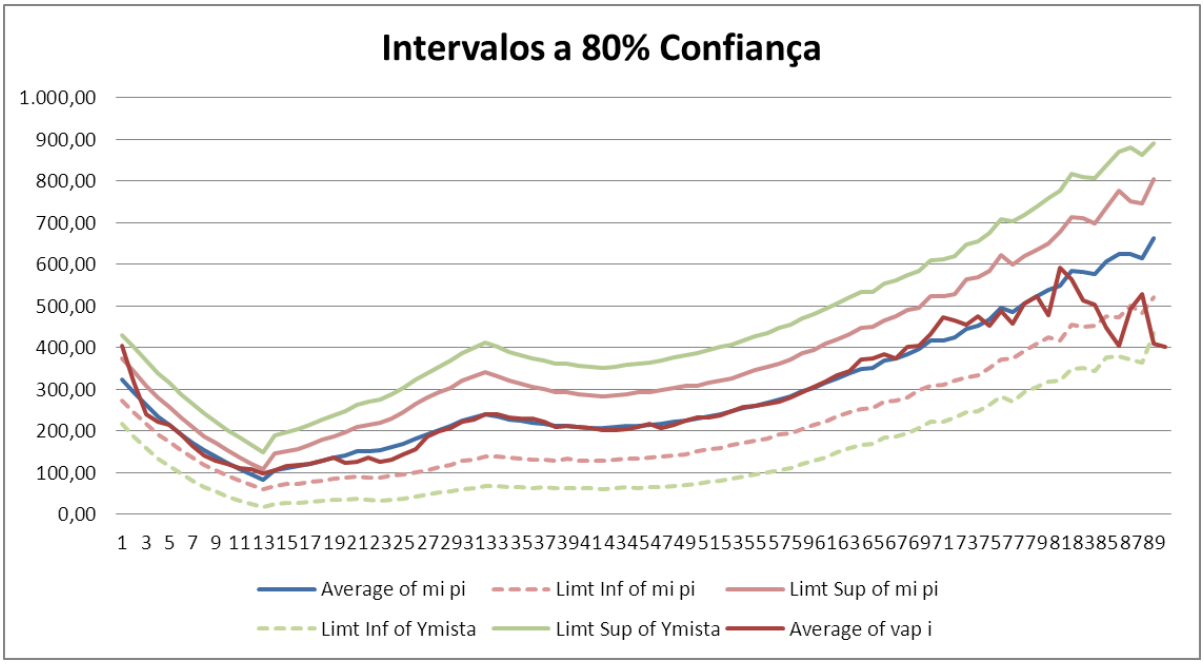
$$P1(idade) = -0{,}000188*idade^2 + 0{,}0141*idade - 0{,}1712$$

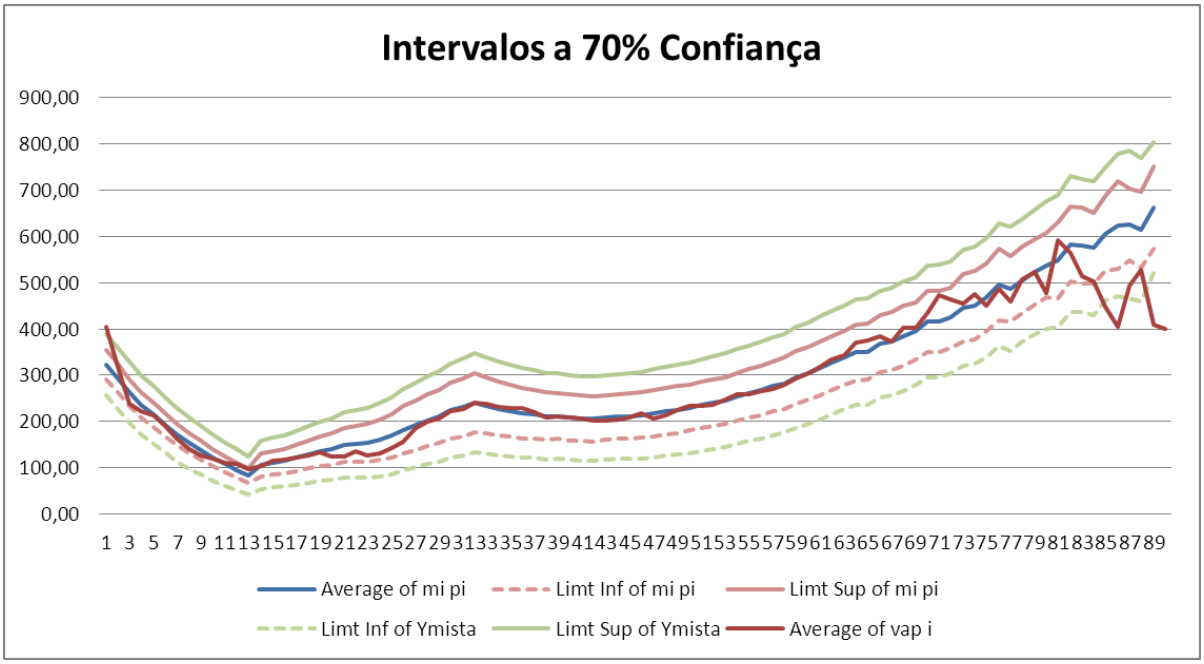
*E acima dos 32 anos o polinómio:*

$$P2(idade) = 0{,}00033*idade^2 - 0{,}03172*idade + 0{,}76578$$

*Utilizando as funções residentes em Excel*

## Anexo 3:


Intervalos a 80% Confiança
1.000,00
900,00
800,00
700,00
600,00
500,00
400,00
300,00
200,00
100,00
0,00
1 3 5 7 9 11131517192123252729313335373941434547495153555759616365676971737577798183858789
Average of mi pi
Limt Inf of mi pi
Limt Sup of mi pi
Limt Inf of Ymista
Limt Sup of Ymista
Average of vap i


Intervalos a 70% Confiança
900,00
800,00
700,00
600,00
500,00
400,00
300,00
200,00
100,00
0,00
1 3 5 7 9 11131517192123252729313335373941434547495153555759616365676971737577798183858789
Average of mi pi
Limt Inf of mi pi
Limt Sup of mi pi
Limt Inf of Ymista
Limt Sup of Ymista
Average of vap i

## Anexo 4:

*Códigos atribuídos às variáveis que foram alvo da Análise de Clusters:*

*Tipo de Produto:*

| Código | Tipo de Produto |
|---|---|
| 1 | CGD |
| 2 | CTT |
| 3 | FM |
| 4 | GRANDES CLIENTES |
| 5 | IB |
| 6 | OUTROS CLIENTES |
| 7 | PMEs |
| 8 | PROTOCOLO |

*Parentesco:*

| Código | Parentesco |
|---|---|
| 1 | Ascendente |
| 2 | Pai ou Mãe |
| 3 | Titular |
| 4 | Conjuge |
| 5 | Outras Relações |
| 6 | Outros Familiares |
| 7 | Filho Adulto |
| 8 | Filho |
| 9 | Neto |

Seguro De Saúde: Custos De Ambulatório - Modelização Linear Generalizada

*Distrito:*

| Código | Distrito1 |
|---|---|
| 1 | AVEIRO |
| 2 | BEJA |
| 3 | BRAGA |
| 4 | BRAGANÇA |
| 5 | CASTELO BRANCO |
| 6 | COIMBRA |
| 7 | ÉVORA |
| 8 | FARO |
| 9 | GUARDA |
| 10 | ILHA DA MADEIRA |
| 11 | ILHA DAS FLORES |
| 12 | ILHA DE PORTO SANTO |
| 13 | ILHA DE SANTA MARIA |
| 14 | ILHA DE SÃO JORGE |
| 15 | ILHA DE SÃO MIGUEL |
| 16 | ILHA DO FAIAL |
| 17 | ILHA DO PICO |
| 18 | ILHA TERCEIRA |
| 19 | LEIRIA |
| 20 | LISBOA |
| 21 | PORTALEGRE |
| 22 | PORTO |
| 23 | SANTARÉM |
| 24 | SETÚBAL |
| 25 | VIANA DO CASTELO |
| 26 | VILA REAL |
| 27 | VISEU |

*Zona Multicare:*

| Código | Zona Multicare |
|---|---|
| 1 | AÇORES |
| 2 | ALGARVE |
| 3 | ALTO ALENTEJO |
| 4 | ALTO MINHO |
| 5 | BAIXO ALENTEJO |
| 6 | BAIXO MINHO |
| 7 | BEIRA INTERIOR NORTE |
| 8 | BEIRA INTERIOR SUL |
| 9 | BEIRA LITORAL CENTRO |
| 10 | BEIRA LITORAL NORTE |
| 11 | BEIRA LITORAL SUL |
| 12 | GRANDE LISBOA |
| 13 | GRANDE PORTO |
| 15 | MADEIRA |
| 16 | MARÃO |
| 17 | MARGEM SUL |
| 18 | OESTE |
| 19 | RIBATEJO |
| 20 | TRÁS-OS-MONTES |

# 6. Bibliografia Consultada

Probabilidades e Estatística:

- Pestana, D. D. e Velosa, S. (2010), Introdução à Probabilidade e à Estatística, 4ª ed., Edição Calouste Gulbenkian.

Modelos Lineares Generalizados:

- Amaral Turkman, M.A. e Silva, G. (2000). Modelos Lineares Generalizados – da Teoria à Prática, Edições SPE, Lisboa.
- Casella, G., Berger, RL. (2002), Statistical Inference,2ª Edição, Pacific Grove, CA: Duzbury Press. Pág. 591-596
- Hosmer, D. e Lemeshow, S. (2000), Applied Logistic Regression, 2ª Edição, New York, New York, USA: A Wiley-Interscience Publication, John Wiley & Sons Inc.
- Gomes, João, 2011, Regressão Binária: O Modelo Logístico, DEIO, FCUL.
- Andreozzi, Valeska, Apontamentos da cadeira Modelos Lineares Generalizados, 2011, DEIO, FCUL.
- Mendes Leal, Margarida, Apontamentos Amostragem e Análise de Dados (Ana´lise de Dados Multivariados), 2011, DEIO, FCUL.

Estudos Aplicados:

- Gravelle, H., Dusheiko, M. et al., Modelling Individual Patiente Hospital Expenditure for General Practice Budgets, 2011, The University of York – Centre For Health Economics, http://www.york.ac.uk/media/che/documents/papers/researchpapers/CHERP73_Modelling_Individual_Patient_Hospital_Expenditure_for_GP_Budgets.pdf.
- Maia, A. C., Andrade, M. V., Chein, F., Estudo Longitudinal do Efeito da Idade e Tempo até à Morte em Gastos com Saúde, 2012, Brasil, http://reap.org.br/wp-content/uploads/2012/05/037-Estudo-Longitudinal-dos-Efeitos-do-Gasto.pdf.
- Agranonik, M., Equações de estimação Generalizada (GEE): Aplicação em Estudo sobre Mortalidade Neonatal em Gemelares de Porto Alegre, RS (1995-2007), 2009, Brasil, http://www.lume.ufrgs.br/bitstream/handle/10183/19081/000735185.pdf?sequence=1.
- Anirban Basu, PhD, Modelling Healthcare Expenditures, 2010, The University of Chicago, http://www.hsrd.research.va.gov/for_researchers/cyber_seminars/archives/hmcs-061610.pdf.